



3 1761 10374365 4

Government  
Publications

---

# Survey Methodology

---

Catalogue No. 12-001-XPB

A journal  
published by  
Statistics Canada

June 2008

•

Volume 34

•

Number 1



Statistics  
Canada

Statistique  
Canada

Canada

## How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1-800-263-1136).

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at [www.statcan.ca](http://www.statcan.ca), e-mail us at [infostats@statcan.ca](mailto:infostats@statcan.ca), or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

### Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

### Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

## To access and order this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website at [www.statcan.ca](http://www.statcan.ca) and select "Publications"> "Free Internet publications."

This product, Catalogue no. 12-001-X, is also available as a standard printed publication at a price of CAN\$30.00 per issue and CAN\$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	Annual subscription
United States	CAN\$6.00	CAN\$12.00
Other countries	CAN\$10.00	CAN\$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

- Telephone (Canada and United States) 1-800-267-6677
- Fax (Canada and United States) 1-877-287-4369
- E-mail [infostats@statcan.ca](mailto:infostats@statcan.ca)
- Mail Statistics Canada  
Finance  
R.H. Coats Bldg., 6th Floor  
150 Tunney's Pasture Driveway  
Ottawa, Ontario K1A 0T6
- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.ca](http://www.statcan.ca) under "About us" > "Providing services to Canadians."



---

# Survey Methodology

---

A journal  
published by  
Statistics Canada

June 2008 • Volume 34 • Number 1

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2008

All rights reserved. This product cannot be reproduced and/or transmitted to any person or organization outside of the licensee's organization. Reasonable rights of use of the content of this product are granted solely for personal, corporate or public policy research, or for educational purposes.

This permission includes the use of the content in analyses and the reporting of results and conclusions, including the citation of limited amounts of supporting data extracted from this product. These materials are solely for non-commercial purposes. In such cases, the source of the data must be acknowledged as follows:

Source (or "Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, users shall seek prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

June 2008

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics  
Canada

Statistique  
Canada

Canada

# SURVEY METHODOLOGY

## A Journal Published by Statistics Canada

*Survey Methodology* is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

### MANAGEMENT BOARD

**Chairman** D. Royce

**Past Chairmen** G.J. Brackstone  
R. Platek

**Members** J. Gambino  
R. Jones  
J. Kovar  
H. Mantel  
E. Rancourt

### EDITORIAL BOARD

**Editor** J. Kovar, *Statistics Canada*  
**Deputy Editor** H. Mantel, *Statistics Canada*

**Past Editor** M.P. Singh

#### Associate Editors

J.M. Brick, *Westat Inc.*  
P. Cantwell, *U.S. Bureau of the Census*  
J.L. Eltinge, *U.S. Bureau of Labor Statistics*  
W.A. Fuller, *Iowa State University*  
J. Gambino, *Statistics Canada*  
M.A. Hidioglou, *Statistics Canada*  
D. Judkins, *Westat Inc.*  
P. Kott, *National Agricultural Statistics Service*  
P. Lahiri, *JPSM, University of Maryland*  
P. Lavallée, *Statistics Canada*  
G. Nathan, *Hebrew University*  
J. Opsomer, *Colorado State University*  
D. Pfeffermann, *Hebrew University*  
N.G.N. Prasad, *University of Alberta*  
J.N.K. Rao, *Carleton University*

T.J. Rao, *Indian Statistical Institute*  
J. Reiter, *Duke University*  
L.-P. Rivest, *Université Laval*  
N. Schenker, *National Center for Health Statistics*  
F.J. Scheuren, *National Opinion Research Center*  
P. do N. Silva, *University of Southampton*  
E. Stasny, *Ohio State University*  
D. Steel, *University of Wollongong*  
L. Stokes, *Southern Methodist University*  
M. Thompson, *University of Waterloo*  
Y. Tillé, *Université de Neuchâtel*  
V.J. Verma, *Università degli Studi di Siena*  
K.M. Wolter, *Iowa State University*  
C. Wu, *University of Waterloo*  
A. Zaslavsky, *Harvard University*

**Assistant Editors** J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer and W. Yung, *Statistics Canada*

### EDITORIAL POLICY

*Survey Methodology* publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

#### Submission of Manuscripts

*Survey Methodology* is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the Journal.

#### Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN \$58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN \$12 (\$6 × 2 issues); Other Countries, CDN \$20 (\$10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: [www.statcan.ca](http://www.statcan.ca).



**Survey Methodology**  
A Journal Published by Statistics Canada  
Volume 34, Number 1, June 2008

**Contents**

In This Issue.....	1
 <b>Regular Papers</b>	
Mary E. Thompson and Changbao Wu Simulation-based randomized systematic PPS sampling under substitution of units .....	3
Mahmoud Torabi and J.N.K. Rao Small area estimation under a two-level model .....	11
Yong You An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada .....	19
Junyuan Wang, Wayne A. Fuller and Yongming Qu Small area estimation under a restriction.....	29
Bal gobin Nandram and Jai Won Choi A Bayesian allocation of undecided voters .....	37
Radu Lazar, Glen Meeden and David Nelson A noninformative Bayesian approach to finite population sampling using auxiliary variables .....	51
Alan M. Zaslavsky, Hui Zheng and John Adams Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates.....	65
Yan Li Generalized regression estimators of a finite population total using the Box-Cox technique .....	79
Cédric Béguin and Beat Hulliger The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data .....	91
Annette Jäckle and Peter Lynn Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias.....	105
Leyla Mohadjer and Lester R. Curtin Balancing sample design goals for the National Health and Nutrition Examination Survey .....	119

$\infty$



## In This Issue

This issue of *Survey Methodology* includes papers on a variety of methodology topics.

In the first paper, Thompson and Wu consider the problem of obtaining inclusion probabilities, for the derivation of sampling weights, when modifications or compromises to the original sample design have been made due to practical constraints or limitations. The problem was motivated by the International Tobacco Control Policy Evaluation Survey of China which used a multi-stage unequal probability design for the selection of adult smokers and nonsmokers from seven cities. Due to refusal to participate by some districts, substitution units had to be selected after the original sample was selected. This substitution made it very difficult to calculate first order inclusion probabilities and practically impossible to calculate second order probabilities. In the paper the authors demonstrate, both theoretically and empirically, that the first and second order inclusions probabilities can be accurately estimated through Monte Carlo simulations.

Torabi and Rao derive the mean squared error (MSE) of a proposed new generalized regression estimator (GREG) of a small area mean under a two-level model and provide both theoretical and empirical comparisons between the new GREG and the best linear unbiased prediction estimator in terms of relative efficiency.

The paper by You discusses various cross-sectional and time series small area models for unemployment rate estimation for Canadian sub-provincial areas. In particular he considers an integrated non-linear mixed effects model under the hierarchical Bayes (HB) framework. An HB approach with the Gibbs sampling method is used to obtain estimates of posterior means and posterior variances of small area unemployment rates. The proposed HB model leads to reliable model-based estimates in terms of CV reduction. You also analyses the proposed model fitness and compares the model-based estimates to direct estimates.

The paper by Wang, Fuller and Qu studies small area estimation under a restriction. The authors study the impact of different augmented models in terms of MSE of the EBLUP. They consider small area models augmented with one additional explanatory variable for which the usual small area predictors achieve a self-calibrated property. They then consider small area models augmented with an added auxiliary variable that is a function of area size to reduce the bias when an incorrect model is used for prediction.

Nandram and Choi present an interesting approach to allocating undecided voters in surveys conducted prior to an election. Data from election polls are typically presented in two-way categorical tables with many polls taken before the actual election. They present the construction and analysis of a time-dependent nonignorable nonresponse model using Bayesian methods. They compare their model to extended versions (to include time) of ignorable and nonignorable nonresponse models introduced by Nandram, Cox and Choi (Survey Methodology, 2005). They also construct a new parameter to help predict the winner. The approach is illustrated using polling data from the 1998 race for governor of Ohio.

In their paper, Lazar, Meeden and Nelson develop a Bayesian approach to finite population sampling, through the use of a Polya posterior, when prior information is available in the form of partial knowledge about an auxiliary variable. The authors introduce the constrained Polya estimator and show that it has similarities with the generalized regression estimator under simple random sampling. However, their estimator does not require specification of a linear model. It is also related to empirical likelihood methods. Examples are used to illustrate the theory.

Zaslavsky, Zheng and Adams consider optimal sampling rates in element-sampling designs when the anticipated analysis is a survey-weighted linear regression and the estimands of interest are linear combinations of regression coefficients from one or more models. Methods are first developed assuming that exact design information is available in the sampling frame and then generalized to situations in which some design variables are available only as aggregates for groups of potential subjects, or from inaccurate or old data. Potential applications include estimation of means for several sets of overlapping domains, estimation for subpopulations such as minority races by disproportionate sampling of geographic areas, and studies in which characteristics available in sampling frames are measured with error.

The paper by Li explores the problem of estimating a finite population total using a nonlinear generalized regression estimator. The Box-Cox technique along with pseudo maximum likelihood estimation is used to obtain data-driven predictions. The author shows that the resulting regression estimator is design-consistent. Its performance is also evaluated through a simulation study.

Béguin and Hulliger extend the BACON algorithm to handle incomplete survey data. The BACON algorithm was developed to identify multivariate outliers using Mahalanobis distance. In the presence of missing values, the EM algorithm can be considered to estimate the covariance matrix at each iteration step of the BACON algorithm. The authors modify the EM algorithm to handle finite population sampling, which they call the EEM (Estimated Expectation Maximization) algorithm, and combine this algorithm with the BACON algorithm. This leads to the proposed BACON-EEM algorithm. It is then applied to two datasets and compared with alternative methods.

The paper by Jäckle and Lynn provides an empirical assessment of the effects of continued incentive payments on attrition, nonresponse bias and item nonresponse, and whether these effects change across waves of a multi-mode panel survey of young people in the UK. They test several hypotheses about the effects of incentives. They conclude that respondent incentives are an effective means of maintaining sample sizes of a panel, thus ensuring its value in terms of efficiency, especially for subgroup analyses. However, they also found that incentives had no effect on attrition bias.

Finally, Mohadjer and Curtin discuss challenges in designing and implementing a sample selection process that satisfies the goals of the National Health and Nutrition Examination Survey (NHANES). They describe how the sample design for NHANES must balance the requirement for efficient subdomain samples with the need for an efficient workload for the interview and examination staff at the Mobile Examination Centres (MEC), while keeping response rates as high as possible and costs down. The article elaborates on a number of unique features of the NHANES design and concludes with a brief summary of what has been achieved and some of the challenges facing future NHANES designs.

Harold Mantel, Deputy Editor



# Simulation-based randomized systematic PPS sampling under substitution of units

Mary E. Thompson and Changbao Wu<sup>1</sup>

## Abstract

The International Tobacco Control (ITC) Policy Evaluation Survey of China uses a multi-stage unequal probability sampling design with upper level clusters selected by the randomized systematic PPS sampling method. A difficulty arises in the execution of the survey: several selected upper level clusters refuse to participate in the survey and have to be replaced by substitute units, selected from units not included in the initial sample and once again using the randomized systematic PPS sampling method. Under such a scenario the first order inclusion probabilities of the final selected units are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In this paper we develop a simulation-based approach for computing the first and the second order inclusion probabilities when direct calculation is prohibitive or impossible. The efficiency and feasibility of the proposed approach are demonstrated through both theoretical considerations and numerical examples. Several R/S-PLUS functions and codes for the proposed procedure are included. The approach can be extended to handle more complex refusal/substitution scenarios one may encounter in practice.

**Key Words:** Inclusion probability; Horvitz-Thompson estimator; Rao-Sampford method; Relative bias; Unequal probability sampling without replacement.

## 1. Introduction

Construction of survey weights is the first critical step in analyzing complex survey data. It starts with the calculation of the first order inclusion probabilities, which is often straightforward if the original sampling design is well executed without any alterations and/or modifications. For instance, if the sample units are selected with inclusion probability ( $\pi$ ) proportional to size (PPS or  $\pi ps$ ), then the inclusion probabilities are readily available from a simple re-scaling of the size variable. Among existing unequal probability without replacement PPS sampling procedures which are applicable for arbitrary fixed sample sizes, the randomized systematic PPS sampling method is the simplest one to implement. The procedure was first described in Goodman and Kish (1950) as a controlled selection method, and was refined by Hartley and Rao (1962) who studied the important and yet difficult problem of how to compute the second order inclusion probabilities. Let  $x_i, i = 1, 2, \dots, N$  be the values of the known size variable, where  $N$  is the total number of units in the population. Let  $z_i = x_i / X$  where  $X = \sum_{i=1}^N x_i$  and assume  $nz_i < 1$  for all  $i$ . The randomized systematic PPS sampling procedure is as follows: Arrange the  $N$  population units in a random order and let  $A_0 = 0$  and  $A_j = \sum_{i=1}^j (nz_i)$  be the cumulative totals of  $nz_i$  in that order so that  $0 = A_0 < A_1 < \dots < A_N = n$ . Let  $u$  be a uniform random number over  $[0, 1]$ . The  $n$  units to be included in the sample are those with indices  $j$  satisfying  $A_{j-1} \leq u + k < A_j$  for  $k = 0, 1, \dots, n-1$ . Let  $s$  be the set of  $n$  sampled units and

$\pi_i = P(i \in s)$  be the first order inclusion probabilities. The randomized systematic PPS sampling procedure satisfies the condition

$$\pi_i = nz_i, \quad i = 1, 2, \dots, N. \quad (1.1)$$

Several other without replacement sampling procedures which satisfy (1.1) for an arbitrary fixed sample size  $n$  were also proposed in the literature, including the well-known Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967) and those of Chao (1982), Chen, Dempster and Liu (1994), Tillé (1996) and Deville and Tillé (1998), among others.

The extensive research work on PPS sampling methods was largely stimulated by the Horvitz-Thompson (HT) estimator  $\hat{T} = \sum_{i \in s} y_i / \pi_i$  for the population total  $T = \sum_{i=1}^N y_i$  of a study variable  $y$ . The HT estimator is extremely efficient when  $y$  is highly correlated with the size variable  $x$  and the sampling procedure satisfies (1.1). It is the unique design unbiased estimator among the class of linear estimators  $\sum_{i \in s} w_i y_i$  for  $T$  if the weights  $w_i$  depend only on  $i$ .

While a PPS sampling procedure can be desirable from a theoretical point of view, it is often difficult and/or sometimes impossible to execute due to practical constraints and limitations. Certain modifications and compromises will have to be made. The modified design, however, will no longer satisfy condition (1.1). Direct calculation of the final inclusion probabilities often becomes difficult or even impossible. Among common problems arising from survey practice which require alteration of the original sampling

1. Mary E. Thompson, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: methomps@uwaterloo.ca; Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo. E-mail: cbwu@uwaterloo.ca.

design, units refusal and substitution of units are the most frequently encountered ones. The scenario is well illustrated by the following example.

The International Tobacco Control (ITC) Policy Evaluation Survey of China (ITC China Survey) uses a multi-stage unequal probability sampling design for the selection of adult smokers and nonsmokers from seven cities. Each city has a natural hierarchical administrative structure

City → Street District → Residential Block → Household → Individual

which was conveniently integrated into the sampling design. At the upper levels, the randomized systematic PPS sampling method is used to select ten street districts from each city, with probability proportional to the population size of the district, and then two residential blocks are selected within each selected district, again using the randomized systematic PPS sampling method, with probability proportional to the population size of the block. Households and individuals within households are further selected, using a modified simple random sampling method. The original plan was to select 40 adult smokers and 10 adult nonsmokers from each of the 20 residential blocks, making the final sample with 800 smokers and 200 nonsmokers for each city.

A difficulty, however, arises in the execution of the survey: several selected upper level clusters (first Street Districts and then Residential Blocks) have refused to participate in the survey, due to time conflict with other activities or unavailability of human resources. These refusing clusters have to be replaced by substitute units, selected from units not included in the initial sample; one possibility is to use once again the randomized systematic PPS sampling method, to achieve the targeted overall sample size.

Under multi-stage sampling designs such as the one used for the ITC China survey, first order inclusion probabilities for individuals selected in the final sample can be calculated by multiplying the inclusion probabilities of units at different stages. When the randomized systematic PPS sampling method is modified due to substitution of units at a certain stage, the condition (1.1) no longer holds for the final sample at that stage. The first order inclusion probabilities under such a scenario are very difficult to calculate and the second order inclusion probabilities become virtually intractable. In Appendix A, we provide a method of direct calculation (5.2) for the  $\pi_i$  when both the initial and the substitute samples are selected using the randomized systematic PPS sampling, assuming random refusal from the initial sample and no refusal from the substitute sample. The expression is valid conditional on the number of refusals and the population order used (after randomization)

for the selection of the initial sample. It is apparent that even under such restrictive conditions and assumptions, the expression itself becomes computationally unfriendly with a not-so-large sample size.

In this paper we demonstrate, through both theory and numerical examples, that the first and the second order inclusion probabilities can be accurately estimated through Monte Carlo simulations when complete design information is available. Our numerical examples are motivated by the ITC China survey for which the randomized systematic PPS sampling serves as a baseline method but our theoretical results and the general methodology apply to other unequal probability without replacement sampling procedures as well. Section 2 presents results on the accuracy of simulation based methods. Numerical examples and comparisons are given in Section 3. Several R/S-PLUS functions and codes for the proposed procedure, originally developed for the ITC China survey, are included in Appendix C. Some additional remarks are given in Section 4.

## 2. Properties of simulation-based methods

When calculation of exact inclusion probabilities is impossible or prohibitive but complete design information is available, Monte Carlo simulation methods can easily be used to obtain estimates of the inclusion probabilities. Denote the completely specified probability sampling design by  $p$ . The simulation-based method is straightforward: select  $K$  independent samples, all following the same sampling design  $p$ ; let  $M_i$  be the number of samples which include unit  $i$ . Then the first order inclusion probability  $\pi_i = P(i \in s)$  can be estimated by  $\pi_i^* = M_i / K$ . For a particular  $i$ , the  $M_i$  follows a binomial distribution and the  $\pi_i^*$  satisfies  $E(\pi_i^*) = \pi_i$  and  $\text{Var}(\pi_i^*) \leq (4K)^{-1}$ . Suppose for instance that we can afford to take  $K$  as big as  $25 \times 10^6$ , then  $P(|\pi_i^* - \pi_i| < 0.001) \geq 0.99$  for any given  $\pi_i$ .

A more relevant measure of the accuracy of simulation-based methods is the performance of the Horvitz-Thompson estimator using the simulated inclusion probabilities. Let  $\hat{T} = \sum_{i \in s} y_i / \pi_i$  and  $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$ . For a given sample, the relative bias of using  $\tilde{T}$  in place of  $\hat{T}$  is defined as  $(\hat{T} - \tilde{T}) / \hat{T}$ . Without loss of generality, we assume  $y_i \geq 0$  for all  $i$ . It is shown in Appendix B that for any  $\varepsilon > 0$  and the given sample  $s$ ,

$$P\left(\left|\frac{\hat{T} - \tilde{T}}{\hat{T}}\right| \leq \varepsilon\right) \geq 1 - \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n\right). \quad (2.1)$$

Note that  $\sum_{i \in s} (1/\pi_i)$  is the Horvitz-Thompson estimator of the population size  $N$ , a practical lower bound for  $P(|\hat{T} - \tilde{T}|/\hat{T} \leq \varepsilon)$  with a small  $\varepsilon$  is given by



$$\Delta = 1 - \frac{2(N-n)}{K\varepsilon^2}. \quad (2.2)$$

If one requires that  $\varepsilon = 0.01$  and  $\Delta = 0.98$ , then for  $N - n = 100$  the (theoretical) number of independent samples required for the simulation is  $K = 10^8$ . Since the lower bound given by (2.1) is conservative, and valid for any response variable, one would expect that a smaller  $K$  with values around  $10^7$  or even  $10^6$  should work well for most practical scenarios where  $N - n \leq 100$ . This is supported by numerical examples presented in Section 3.

Estimation of the second order inclusion probabilities  $\pi_{ij} = P(i, j \in s)$  imposes no additional difficulty except that the total number of simulated samples,  $K$ , required to achieve the same level of relative accuracy as for the first order case is bigger. Let  $M_{ij}$  be the number of simulated samples among the  $K$  independent samples which include both  $i$  and  $j$ . Let  $\pi_{ij}^* = M_{ij} / K$  be the estimate for  $\pi_{ij}$ . Suppose the goal is to estimate a quadratic population quantity

$$Q = \sum_{i=1}^N \sum_{j=1}^N q(y_i, y_j).$$

The Horvitz-Thompson type estimators of  $Q$  using  $\pi_{ij}$  or  $\pi_{ij}^*$  are respectively given by

$$\hat{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}} \text{ and } \tilde{Q} = \sum_{i \in s} \sum_{j \in s} \frac{q(y_i, y_j)}{\pi_{ij}^*}.$$

Following the same argument as that which leads to (2.1), we can show that

$$P\left(\frac{|\hat{Q} - \tilde{Q}|}{\tilde{Q}} \leq \varepsilon\right) \geq 1 - \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left( \sum_{i \in s} \sum_{j \in s} \frac{1}{\pi_{ij}} - n^2 \right). \quad (2.3)$$

Note that  $\sum_{i \in s} \sum_{j \in s} (1/\pi_{ij})$  is a design-unbiased estimator of  $N^2$ , a practical lower bound for  $P(|\hat{Q} - \tilde{Q}|/\tilde{Q} \leq \varepsilon)$  is given by  $1 - 2(N+n)(N-n)/(K\varepsilon^2)$ . Comparing this with  $\Delta$  given by (2.2), it is apparent that we need a much bigger  $K$  to achieve the same lower bound, although in both cases the lower bounds are conservative, and the actual  $K$  required can be smaller. On the other hand, second order inclusion probabilities are used for the estimation of second order parameters such as the population variance or the variance of a linear estimator. The desired estimation accuracy is less critical than that for first order parameters such as the population total or mean, and therefore a number in between  $10^6$  and  $10^7$  for  $K$  should be acceptable for many practical situations.

The most critical issue for simulation-based methods is obviously the feasibility of computational implementation. Among other things, it depends largely on the chosen value of  $K$ , the complexity of the sampling design, and the

computational power available. If  $K = 10^6$  and one would like to have the simulation-based results within ten hours, then it is necessary to take 28 simulated samples for every single second. The randomized systematic PPS sampling is the most efficient unequal probability without replacement sampling procedure in terms of computational implementation. It only involves a simple random ordering and selecting a random starting point. Most other competing procedures involve either rejective methods or complicated sequential selections. It takes much longer to select simulated samples with these methods. A comparison of CPU times for computing the simulated  $\pi_i$  between the randomized systematic PPS sampling and the Rao-Sampford unequal probability sampling design is given in Section 4.

### 3. Numerical examples

The design information used in this section is adapted from the ITC China survey. The number of Street Districts (top level clusters) in each of the seven cities involved in the survey ranges from  $N = 20$  to  $N = 120$ . Within each city  $n = 10$  districts are selected using the randomized systematic PPS sampling method. In the case of refusals, substitute districts are selected from the ones not included in the initial sample, again using the randomized systematic PPS sampling method. For the purpose of illustration we use the design information from the smallest city (*i.e.*,  $N = 20$ ). Additional comments on cases where  $N$  is large are given in Section 4.

#### 3.1 First order inclusion probabilities

We first demonstrate the accuracy of the simulated  $\pi_i$  when the exact values of  $\pi_i$  are known. We then investigate the impact of substitution of units on the final  $\pi_i$  and the performance of the Horvitz-Thompson estimator for a population total using the simulated  $\pi_i$ . The simulated inclusion probabilities under substitution of units are compared to those assuming the modified design is still PPS sampling.

*Example 1.* Simulation-based  $\pi_i^*$  when there is no refusal. In this case the exact values of  $\pi_i$  are given by  $\pi_i = nz_i$ .

(i) Exact values of  $\pi_i$ :

0.5840 0.5547 0.6702 0.5331 0.3085 0.2652 0.3930 0.4180 0.6952 0.3471  
0.5993 0.5393 0.8240 0.6868 0.4469 0.2191 0.4237 0.4180 0.7567 0.3163

(ii) Simulated  $\pi_i^*$ ,  $K = 10^5$ :

0.5828 0.5545 0.6656 0.5339 0.3071 0.2656 0.3929 0.4205 0.6969 0.3474  
0.6009 0.5429 0.8227 0.6865 0.4446 0.2186 0.4215 0.4179 0.7569 0.3194

(iii) Simulated  $\pi_i^*$ ,  $K = 10^6$ :

0.5836 0.5558 0.6701 0.5336 0.3081 0.2654 0.3931 0.4180 0.6950 0.3469  
0.5994 0.5394 0.8242 0.6864 0.4469 0.2186 0.4237 0.4172 0.7569 0.3166

The simulated  $\pi_i^*$  matches  $\pi_i$  to the second decimal point for  $K=10^5$  and to the third for  $K=10^6$  for most cases.

**Example 2.** To assess the performance of the Horvitz-Thompson (HT) estimator for a population total using the true  $\pi_i$  and the simulated  $\pi_i^*$  from Example 1, we generated the response variable from the model  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, \dots, N$ , where  $x_i$  is the size variable and  $\varepsilon_i$  are independent and identically normally distributed with mean 0 and variance  $\sigma^2$ . We considered three populations (three values of  $\sigma^2$ ) where the population correlation coefficients between  $x$  and  $y$  are respectively 0.3, 0.5 and 0.8. For each of the three populations,  $B = 2,000$  repeated samples of size  $n = 10$  were selected using the randomized systematic PPS sampling, and for each sample three HT estimators were computed using the true  $\pi_i$ , the simulated  $\pi_i^*$  with  $K = 10^5$  and the  $\pi_i^*$  with  $K = 10^6$ , respectively. The results, not reported here to save space, showed that all three HT estimators have relative bias less than 0.04% and almost identical mean squared errors.

**Example 3.** When there are refusals in the initial PPS sample and substitute units are selected from units not included in the initial sample using the same PPS sampling procedure, there are two questions of interest: (1) how to compute the inclusion probabilities  $\pi_i$  for the final sample; and (2) to what extent the substitution procedure has altered the original PPS sampling design. We can compute the simulated  $\pi_i^*$  and compare them with  $\tilde{\pi}_i$  obtained by assuming a PPS sampling after the refusing units are removed from the sampling frame. In simulating the  $\pi_i^*$ , we assume for simplicity that there is no possible refusal from any unit outside the initial sample, and hence there is no refusal among the substitute units. The number of replications  $K$  is chosen as  $10^6$  for the simulation. We consider two scenarios where there are three refusing units in the population, and all are among the initial sample of size  $n = 10$ .

- (i) Three large units refuse: Simulated  $\pi_i^*$  (first two rows) versus  $\tilde{\pi}_i$  (last two rows) assuming PPS.

0.7231	0.6981	0.7947	0.6773	0.4354	0.3811	0.5339	0.5619	0.0000	0.4815
0.7363	0.6826	0.0000	0.8070	0.5919	0.3210	0.5678	0.5615	0.0000	0.4441
0.7560	0.7182	0.8677	0.6901	0.3994	0.3434	0.5088	0.5412	0.0000	0.4494
0.7759	0.6983	0.0000	0.8892	0.5786	0.2837	0.5486	0.5412	0.0000	0.4096

- (ii) Three small units refuse: Simulated  $\pi_i^*$  (first two rows) versus  $\tilde{\pi}_i$  (last two rows) assuming PPS.

0.6326	0.6049	0.7167	0.5829	0.0000	0.0000	0.4415	0.4668	0.7406	0.3937
0.6482	0.5901	0.8558	0.7330	0.4965	0.0000	0.4728	0.4664	0.7976	0.3590
0.6343	0.6025	0.7280	0.5790	0.0000	0.0000	0.4268	0.4540	0.7550	0.3770
0.6510	0.5858	0.8949	0.7459	0.4854	0.0000	0.4602	0.4540	0.8218	0.3436

It is apparent that the sizes of the refusing units have dramatic impact on the distribution of the final inclusion probabilities. If one ignores the alteration of the sampling

design due to substitution of units and treats the design as if it is still a PPS sampling, then the inclusion probabilities for large units are inflated and the role of small units is downplayed. This trend is more pronounced when there are large units among the refusals, *i.e.*, case (i) where  $\pi_{i_4}^* = 0.8070$  compared to  $\tilde{\pi}_{i_4} = 0.8897$  and  $\pi_{i_6}^* = 0.3210$  to  $\tilde{\pi}_{i_6} = 0.2837$ .

### 3.2 Second order inclusion probabilities

There have been considerable research activities on the randomized systematic PPS sampling, mainly for obtaining second order inclusion probabilities  $\pi_{ij}$  and variance estimators. Hartley and Rao (1962) derived exact formulas for the  $\pi_{ij}$  when  $n=2$  and  $N=3$  or  $N=4$ ; Connor (1966) extended the results and derived the exact formula for general  $n$  and  $N$ , and the related computational procedure was later implemented in the Fortran language by Hidioglou and Gray (1980). The procedure is quite heavy as evidenced by the 165 lines of Fortran code.

The most intriguing result is probably the asymptotic approximation to  $\pi_{ij}$  derived by Hartley and Rao (1962). In a recent paper Kott (2005) showed that the variance estimator of a Horvitz-Thompson estimator based on the Hartley-Rao approximation not only performs well under the design-based framework but also has good model-based properties. The Hartley-Rao approximation was initially derived based on the assumption that  $n$  is fixed and  $N$  is large and is correct to the order of  $O(N^{-4})$  (Hartley and Rao 1962: Equation (5.15) on page 369). In a private conversation with J.N.K. Rao during the 23<sup>rd</sup> International Methodological Symposium of Statistics Canada, he pointed out that the approximation is still valid even if  $n$  is large, as long as  $n/N$  is small. For cases where  $N$  is not large and/or  $n/N$  is not small, such as the ITC China survey example considered here, the goodness of the Hartley-Rao approximation has not been documented.

When the randomized systematic PPS sampling procedure is altered due to substitution of units, it is virtually impossible to derive the second order inclusion probabilities or some sort of approximations. With the simulation-based approach, however, it remains straightforward to obtain very reliable estimates of the  $\pi_{ij}$  through a large number of simulated samples, given that the altered sampling procedure is completely specified. In what follows we examine the performance of variance estimators using the simulated  $\pi_{ij}^*$  when there is no alteration to the randomized systematic PPS sampling procedure. In this case  $\pi_i = n\pi_i^*$  and the Hartley-Rao approximation  $\tilde{\pi}_{ij}$  to  $\pi_{ij}$  can also take part in the comparison.

**Example 4.** We first compare  $\pi_{ij}^*$  to  $\tilde{\pi}_{ij}$  for each of the individual entries. To save space, we only present the results for  $i = 1, \dots, 5$  and  $j = 1, \dots, 10$ , which are sufficient to



show the general picture. The Hartley-Rao approximation  $\tilde{\pi}_{ij}$  is very close to the simulated  $\pi_{ij}^*$ , matching to the second decimal point for the majority of the entries. This is clearly an interesting observation given that  $N = 20$  and  $n = 10$ .

(i) Simulated  $\pi_{ij}^*$ ,  $K = 10^6$ :

0.0000	0.3121	0.3821	0.2975	0.1669	0.1442	0.2116	0.2249	0.3975	0.1873
0.3121	0.0000	0.3623	0.2816	0.1590	0.1372	0.2025	0.2141	0.3766	0.1784
0.3821	0.3623	0.0000	0.3469	0.1899	0.1640	0.2483	0.2659	0.4586	0.2153
0.2975	0.2816	0.3469	0.0000	0.1523	0.1312	0.1938	0.2061	0.3606	0.1717
0.1669	0.1590	0.1899	0.1523	0.0000	0.0742	0.1124	0.1197	0.1968	0.0988

(ii) Hartley-Rao approximation  $\tilde{\pi}_{ij}$ :

0.0000	0.3079	0.3769	0.2952	0.1668	0.1427	0.2143	0.2286	0.3921	0.1884
0.3079	0.0000	0.3569	0.2795	0.1579	0.1351	0.2029	0.2164	0.3712	0.1784
0.3769	0.3569	0.0000	0.3421	0.1932	0.1654	0.2484	0.2649	0.4544	0.2183
0.2952	0.2795	0.3421	0.0000	0.1514	0.1296	0.1946	0.2075	0.3559	0.1710
0.1668	0.1579	0.1932	0.1514	0.0000	0.0732	0.1099	0.1172	0.2010	0.0966

*Example 5.* For second order inclusion probabilities the main focus is on variance estimation. With fixed sample size, an unbiased variance estimator for the Horvitz-Thompson estimator  $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$  is given by the well-known Yates-Grundy format,

$$v(\hat{Y}_{HT}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left( \frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2. \tag{3.1}$$

We consider the three synthetic populations described in *Example 2*. The true variance  $V = \text{Var}(\hat{Y}_{HT})$  is obtained through simulation using  $B = 10^5$  simulated samples and is computed as  $B^{-1} \sum_{b=1}^B (\hat{Y}_b - Y)^2$ , where  $Y$  is the true population total and  $\hat{Y}_b$  is the Horvitz-Thompson estimator of  $Y$  computed from the  $b^{\text{th}}$  simulated sample. Three variance estimators in the form of (3.1), denoted respectively by  $v_1, v_2$  and  $v_3$ , are examined, with the  $\pi_{ij}$  in (3.1) being respectively replaced by the Hartley-Rao approximation  $\tilde{\pi}_{ij}$ , the simulated  $\pi_{ij}^*$  for  $K = 10^5$  and the  $\pi_{ij}^*$  for  $K = 10^6$ . The performance of these estimators is measured through the simulated relative bias  $\text{RB} = B^{-1} \sum_{b=1}^B (v^{(b)} - V) / V$  and the simulated instability  $\text{INST} = \{B^{-1} \sum_{b=1}^B (v^{(b)} - V)^2\}^{1/2} / V$ , where  $v^{(b)}$  is the variance estimate computed from the  $b^{\text{th}}$  sample, using another set of  $B = 10^5$  independent samples. The results are summarized in Table 1 below. The three populations are indicated by the correlation coefficient  $\rho$  between  $y$  and  $x$ .

Table 1 Relative bias and instability of variance estimators

Population	RB(%)			INST		
	$v_1$	$v_2$	$v_3$	$v_1$	$v_2$	$v_3$
$\rho = 0.30$	6.1%	1.4%	-0.3%	0.66	0.65	0.65
$\rho = 0.50$	4.3%	2.5%	-1.1%	0.42	0.44	0.42
$\rho = 0.80$	2.6%	1.2%	-0.2%	0.61	0.60	0.60

In terms of relative bias, all three variance estimators are acceptable, with the one ( $v_1$ ) based on the Hartley-Rao approximation  $\tilde{\pi}_{ij}$  having the largest bias. For variance

estimators using the simulated  $\pi_{ij}^*$ , increasing the value of  $K$  from  $10^5$  (i.e.,  $v_2$ ) to  $10^6$  (i.e.,  $v_3$ ) makes the bias to be negligible, although the one with  $K = 10^5$  is clearly acceptable in practice. All three versions of the variance estimator have similar measures in terms of instability.

4. Some additional remarks

In theory, the simulation-based method for computing inclusion probabilities is applicable to any sampling design, as long as the complete design information is available. It is an effective approach to handling more complex substitution scenarios or other types of modifications to the original design. In the ITC China survey, one of the refusing units has to be substituted by a unit from a particular region of the city due to workload constraints and field work restrictions. In a Canadian national survey of youth, there were second and third round refusing units (schools) and hence substitute units before achieving the targeted sample size. As pointed out by an Associate Editor, a similar situation was also reported in the 57<sup>th</sup> Round of the National Sample Survey Organization, Government of India ([www.mospi.gov.in](http://www.mospi.gov.in)) where a modification was made to the circular systematic sampling with probability proportional to size in order to select two distinct sub-samples. Gray (1973) described a method on increasing the sample size (number of psu's) when the initial sample was selected by the randomized systematic PPS method. Calculation of second order inclusion probabilities under the proposed procedure is difficult even for a very small sample size. In all these cases analytic solutions to the inclusion probabilities are either difficult to use or not available but the simulation-based approach can be applied without any extra difficulty.

The recent paper by Fattorini (2006) discussed the use of the simulation-based method for spatial sampling where the units are selected sequentially. When a PPS sampling design is altered due to one or more rounds of substitution of units, the modified design can also be viewed as sequential. Our theoretical results on the accuracy of simulation-based methods, however, are different from those of Fattorini. We have used a conditional argument and proposed to assess the performance of the estimator using the simulated inclusion probabilities for a given sample, which is of interest for practical applications.

The central issue related to simulation-based methods is the feasibility of computational implementation. The randomized systematic PPS sampling has a major advantage in computational efficiency. The Rao-Sampford unequal probability sampling method (Rao 1965; Sampford 1967), for instance, is another popular PPS sampling procedure. It has several desirable features such as closed form expressions for the second order inclusion probabilities and

is more efficient than the randomized systematic PPS sampling (Asok and Sukhatme 1976). The following is a comparison of CPU times between the randomized systematic PPS sampling and the Rao-Sampford PPS sampling for simulating the first order inclusion probabilities. The sample size is fixed at  $n=10$  and the number of simulated samples is  $K=10^6$ . The results are obtained using R on a dual-processor unix machine.

N	Systematic PPS	Rao-Sampford PPS
200	4.7 hours	7.5 hours
100	2.5 hours	5.0 hours
50	1.6 hours	4.4 hours
20	1.2 hours	8.9 hours

It is interesting to note that, although in general the Rao-Sampford procedure takes longer time to obtain the results, it takes much longer for the case of  $N=20$ . This is because the Rao-Sampford method uses a rejective procedure and it usually takes many rejections to arrive at a final sample when the sampling fraction  $n/N$  is large. The randomized systematic PPS sampling, on the other hand, is not affected by this and the simulation-based method can provide results with desired accuracy in a timely fashion for  $N=400$  or even bigger. Several R/S-PLUS functions and major codes for the proposed approach are included in Appendix C and are applicable to other substitution scenarios after minor modifications.

One of the reasons for the use of the randomized systematic PPS sampling in selecting upper level clusters in the ITC China survey is that the final design is self-weighting. An interesting question arises when there are refusals: how to select the substitute units such that the final altered sampling design is still (approximately) self-weighting? In some other circumstances such as rotating samples, this is achievable; see, for instance, Fellegi (1963). How to accomplish this goal with the ITC China survey design is currently under investigation.

### Acknowledgements

This research is partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors also thank the International Tobacco Control (ITC) Policy Evaluation Project and the ITC China Survey Project for assistance and support. The ITC project is supported in part by grants from the National Cancer Institute of the United States (P50 CA11236) Roswell Park Transdisciplinary Tobacco Use Research Center and the Canadian Institutes of Health Research (57897). Funding for the ITC China project is provided by the Ministry of Health and the Ministry of Finance of China.

## Appendix A

### A direct calculation under random refusal

Under the randomized systematic PPS sampling design and assuming random refusal, it is possible in principle to calculate the inclusion probabilities under a substitution rule directly. The starting point is to enumerate all possible initial samples and their probabilities based on the particular population order used to select the initial sample.

Recall that  $A_0 = 0$ ,  $A_j = \sum_{i=1}^j (nz_i)$  and  $A_N = n$ . For a chosen uniform starting value  $u \in [0, 1]$ , unit  $j$  is to be selected if

$$A_{j-1} \leq u + k < A_j \quad (5.1)$$

for some  $k = 0, 1, \dots, n-1$ . Let  $k_j$  be the largest integer less than  $A_j$ , and let the remainder  $e_j$  be given by  $e_j = A_j - k_j$ . Let  $0 < e_{(1)} \leq e_{(2)} \leq \dots \leq e_{(N)}$  be the order statistics of the remainders, and let  $k_{(1)}, \dots, k_{(N)}$  be the corresponding  $k_j$ 's. Note that  $e_{(N)} = 1$ . We could then generate  $N$  possible samples  $s_1, \dots, s_N$  with respective probabilities

$$e_{(1)}, e_{(2)} - e_{(1)}, \dots, e_{(N)} - e_{(N-1)},$$

some of which may be 0. We begin by generating  $s_1$ . From each  $j = 1, \dots, N$ , put  $j$  in  $s_1$  if  $A_{j-1} \leq k < A_j$  for some  $k = 0, 1, \dots, n-1$ , i.e.,  $s_1$  is selected using  $u = 0$  in (5.1). As we move  $u$  from 0 to 1, different possible samples can be identified sequentially. Now given  $s_1, \dots, s_m$ , let  $s_{m+1}$  be the same as  $s_m$  except that the  $(k_{(m)} + 1)^{\text{th}}$  element is advanced by 1. For example, suppose  $n = 4$  and  $s_m = \{1, 3, 6, 9\}$ , and suppose  $k_{(m)} = 0$ , then  $s_{m+1} = \{2, 3, 6, 9\}$ . On the other hand, if  $k_{(m)} = 2$ , then  $s_{m+1} = \{1, 3, 7, 9\}$ . The sample  $s_{m+1}$  will have probability  $e_{(m+1)} - e_{(m)}$ .

By construction,  $\pi_i = nz_i$  for  $i = 1, \dots, N$ . If only first and second order inclusion probabilities are desired, a similar but simpler algorithm can be used to calculate the second order inclusion probabilities directly, conditional on the initial order. However, for applications where the probabilities of all samples are needed, the sample generation algorithm can be carried out. For example, for small populations, it is then also possible to calculate the first order inclusion probabilities when there is refusal and substitution. Suppose we first select a sample of size  $n$  with randomized systematic PPS sampling. Suppose  $n_1$  of these agree to respond and an additional  $n_2 = n - n_1$  are selected, again using randomized systematic PPS sampling, from those units not sampled the first time. Assume for simplicity that refusal in the first sample occurs at random, and that there is no refusal in the second substitute sample. Note that this is a different assumption from the one used in Example 3, where the set of refusals is considered to be non-random. The



inclusion probability for unit  $i$ , conditional on the assumed initial population order, is

$$nz_i \times \frac{n_i}{n} + \sum_{m: i \in s_m} p_i(s_m) \frac{n_2 z_i}{\sum_{j: j \in s_m} z_j}. \quad (5.2)$$

The outer sum is taken over all samples  $s_m$  of size  $n$ , generated according to the procedure described above but without having unit  $i$ , with probabilities  $p_i(s_m) = e_{i(m)} - e_{(m-1)}$ . The inner sum involved in the denominator is taken over all  $j$  not included in  $s_m$  from the outer sum. The unconditional inclusion probability can be obtained by appropriate averaging over all population orders which give distinct values. Clearly this is feasible only when the population is small, or when  $z$  takes a small number of values.

## Appendix B

### Derivation of (2.1)

In this appendix we show that for any  $\varepsilon > 0$  and a given sample  $s$ ,

$$P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \leq \varepsilon\right) \geq 1 - \frac{2(1+\varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n\right),$$

where  $\hat{T} = \sum_{i \in s} y_i / \pi_i$ ,  $\tilde{T} = \sum_{i \in s} y_i / \pi_i^*$ , and  $\pi_i^*$  are the simulated first order inclusion probabilities based on  $K$  independent samples. Noting that  $E(\pi_i^*) = \pi_i$  and  $\text{Var}(\pi_i^*) = \pi_i(1 - \pi_i)/K$ , by Chebyshev's inequality we have  $P(|\pi_i^* - \pi_i| > c) \leq \pi_i(1 - \pi_i)/(Kc^2)$  for any  $c > 0$ . It follows that

$$\begin{aligned} P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i} > \varepsilon\right) &= P(\pi_i^* - \pi_i > \pi_i^* \varepsilon) + P(\pi_i^* - \pi_i < -\pi_i^* \varepsilon) \\ &= P(\pi_i^* - \pi_i > \varepsilon \pi_i / (1 - \varepsilon)) + P(\pi_i^* - \pi_i < -\varepsilon \pi_i / (1 + \varepsilon)) \\ &\leq P(|\pi_i^* - \pi_i| > \varepsilon \pi_i / (1 - \varepsilon)) + P(|\pi_i^* - \pi_i| > \varepsilon \pi_i / (1 + \varepsilon)) \\ &\leq \frac{(1 - \varepsilon)^2 \pi_i (1 - \pi_i)}{K\varepsilon^2 \pi_i^2} + \frac{(1 + \varepsilon)^2 \pi_i (1 - \pi_i)}{K\varepsilon^2 \pi_i^2} \\ &= \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left(\frac{1}{\pi_i} - 1\right). \end{aligned}$$

If  $y_i \geq 0$  for all  $i$ , then

$$|\hat{T} - \tilde{T}| \leq \sum_{i \in s} \frac{y_i}{\pi_i} \frac{|\pi_i^* - \pi_i|}{\pi_i^*} \leq \max_{i \in s} \left\{ \frac{|\pi_i^* - \pi_i|}{\pi_i^*} \right\} \hat{T}.$$

For any  $\varepsilon > 0$  and the given sample  $s$ ,

$$\begin{aligned} P\left(\frac{|\hat{T} - \tilde{T}|}{\hat{T}} \leq \varepsilon\right) &\geq P\left(\max_{i \in s} \left\{ \frac{|\pi_i^* - \pi_i|}{\pi_i^*} \right\} \leq \varepsilon\right) \\ &\geq 1 - \sum_{i \in s} P\left(\frac{|\pi_i^* - \pi_i|}{\pi_i^*} > \varepsilon\right) \\ &\geq 1 - \frac{2(1 + \varepsilon^2)}{K\varepsilon^2} \left(\sum_{i \in s} \frac{1}{\pi_i} - n\right). \end{aligned}$$

## Appendix C

### R/S-PLUS Implementation

C1. An R function for randomized systematic PPS sampling.

The input variables of the function are  $x$ : the population vector of size variable and  $n$ : the sample size. The function `syspps` returns the set of  $n$  selected units.

```
syspps<-function(x,n){
  N<-length(x)
  U<-sample(N,N)
  xx<-x[U]
  z<-rep(0,N)
  for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
  r<-runif(1)
  s<-numeric()
  for(i in 1:N){
    if(z[i]>=r){
      s<-c(s,U[i])
      r<-r+1
    }
  }
  return(s[order(s)])
}
```

C2. An R function for simulating the second order inclusion probabilities.

The input variables of the function are  $x$ : the population vector of size variable and  $s$ : the set of labels of units in the sample. The default sampling procedure is the randomized systematic PPS sampling method and the number of repeated samples is  $K = 10^6$ . The function `piij` returns an  $n \times n$  matrix with the  $(ij)^{\text{th}}$  entry being the simulated  $\pi_{ij}^*$ ,  $i, j \in s$ .

```
piij<-function(x,s){
  N<-length(x)
  n<-length(s)
  p<-matrix(0,n,n)
  for(k in 1:1000000){
    ss<-syspps(x,n)
    for(i in 1:(n-1)){
      for(j in (i+1):n){
        if(min(abs(ss-s[i]))+min(abs(ss-s[j]))==0)
          p[i,j]<-p[i,j]+1
      }
    }
    p<-(p+t(p))/1000000
  }
  return(p)
}
```

### C3. An R function for PPS sampling under substitution of units.

```
sysppssub<-function(x,n,refus) {
  s<-syspps(x,n)
  sub<-numeric()
  for (i in 1:n) {
    if (min(abs(s[i]-refus))==0) sub<-c(sub,i)
  }
  m<-length(sub)
  if (m>0) {
    s<-s[-sub]
    U1<-(1:length(x)) [-c(refus,s)]
    x1<-x[-c(refus,s)]
    s1<-syspps(x1,m)
    s<-c(s,U1[s1])
  }
  return(s[order(s)])
}
```

The default procedure for the selection of the initial sample and the substitute sample is the randomized systematic PPS sampling. The following R function `sysppssub` is used for simulating the inclusion probabilities under substitution of units. The input variables are `x`: the population vector of size variable, `n`: the sample size, and `refus`: the set of refusing units from the initial sample. The function returns a set of units for the final sample.

### C4. R codes for simulating the $\pi_i$ under substitution of units.

```
pi<-rep(0,N)
for(i in 1:1000000){
  s<-sysppssub(x,n,refus)
  for(j in 1:N){
    if (min(abs(s-j))==0) pi[j]<-pi[j]+1
  }
}
pi<-pi/1000000
```

## References

Asok, C., and Sukhatme, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

- Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.
- Chen, X.H., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.
- Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61, 384-390.
- Deville, J.C., and Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.
- Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.
- Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.
- Goodman, R., and Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.
- Gray, G.B. (1973). On increasing the sample size (number of psu's). Technical Memorandum, Statistics Canada.
- Hartley, H.O., and Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.
- Hidiroglou, M.A., and Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.
- Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.
- Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statist. Association*, 3, 173-180.
- Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.
- Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.



# Small area estimation under a two-level model

Mahmoud Torabi and J.N.K. Rao<sup>1</sup>

## Abstract

Lehtonen and Veijanen (1999) proposed a new model-assisted generalized regression (GREG) estimator of a small area mean under a two-level model. They have shown that the proposed estimator performs better than the customary GREG estimator in terms of average absolute relative bias and average median absolute relative error. We derive the mean squared error (MSE) of the new GREG estimator under the two-level model and compare it to the MSE of the best linear unbiased prediction (BLUP) estimator. We also provide empirical results on the relative efficiency of the estimators. We show that the new GREG estimator exhibits better performance relative to the customary GREG estimator in terms of average MSE and average absolute relative error. We also show that, due to borrowing strength from related small areas, the EBLUP estimator exhibits significantly better performance relative to the customary GREG and the new GREG estimators. We provide simulation results under a model-based set-up as well as under a real finite population.

Key Words: BLUP estimator; GREG estimator; Mean squared error; Random effects; Small area means.

## 1. Introduction

Small area estimation has received a lot of attention in recent years due to growing demand for reliable small area statistics. Traditional area-specific direct estimators do not provide adequate precision because sample sizes in small areas are seldom large enough. This makes it necessary to employ indirect estimators that borrow strength from related areas, in particular, model-based indirect estimators. Unit level random effect models, including nested error linear regression models and two-level models, are often used in small area estimation to obtain efficient model-based estimators of small area means. Rao (2003) gives a comprehensive account of model-based small area estimation.

A two-level model is given by

$$y_{ij} = x'_{ij}\beta_i + e_{ij};$$

$$\beta_i = Z_i\beta + v_i, j = 1, \dots, N_i; i = 1, \dots, m \quad (1)$$

where  $N_i$  is the number of units in the  $i^{\text{th}}$  area ( $i = 1, \dots, m$ ),  $y_{ij}$  is the response and  $x_{ij}$  is a  $p \times 1$  vector of unit level covariates attached to the  $j^{\text{th}}$  unit in the  $i^{\text{th}}$  area. Further,  $Z_i$  is a  $p \times q$  matrix of area level covariates,  $\beta$  is a  $q \times 1$  vector of regression parameters,  $v_i$ 's are independent random vectors with mean zero and covariance  $\Sigma_v$ , and  $e_{ij}$ 's are independent random variables with mean zero and variance  $\sigma_e^2$  and independent of  $v_i$ 's. We can express the mean  $\bar{y}_i$  of  $i^{\text{th}}$  area as

$$\bar{Y}_i \approx \mu_i = \bar{X}'_i(Z_i\beta + v_i),$$

assuming  $N_i$  is large, where  $\bar{X}_i$  is the known population mean of  $x_{ij}$  in the  $i^{\text{th}}$  area. The sample values  $\{(y_{ij}, x_{ij}), j = 1, \dots, n_i; i = 1, \dots, m\}$  are assumed to obey the model (1), that is, there is no sample selection bias. The model for the sample is then given by

$$y_{ij} = x'_{ij}(Z_i\beta + v_i) + e_{ij}, j = 1, \dots, n_i; i = 1, \dots, m. \quad (2)$$

In matrix notation, (2) may be written as

$$y_i = X_i(Z_i\beta + v_i) + e_i, i = 1, \dots, m$$

with  $\text{Var}(y_i) = V_i = X_i \Sigma_v X'_i + \sigma_e^2 I_{n_i}$ , where  $y_i$  is a  $n_i \times 1$  vector and  $X_i$  is a  $n_i \times p$  matrix. The two-level model (2) was first introduced by Moura and Holt (1999) in the context of small area estimation. This model effectively integrates the use of unit level and area level covariates into a single model, by modeling the random slopes  $\beta_i$  in (1) in terms of area level covariates  $Z_i$ .

Lehtonen and Veijanen (1999) proposed a model-assisted new generalized regression (GREG) estimator of a small area mean under the two-level model. Lehtonen and Veijanen (1999) showed that the new GREG estimator based on model (1) performs better than the customary GREG estimator based on a model with fixed  $\beta_i = Z_i\beta$ . Moura and Holt (1999) obtained the best linear unbiased prediction (BLUP) estimator of the small area mean  $\mu_i$  and its MSE under the two-level model (2); see Section 2. Lehtonen, Särndal, and Veijanen (2003) studied the effect of model choice on different types of estimators (synthetic, GREG, and composite) of small area means.

In Section 3, we first derive the mean squared error (MSE) of the new GREG estimator and the customary GREG estimator (Section 2) under the two-level model (2), assuming known model parameters. We then compare the MSE of the GREG, new GREG and BLUP estimators, and obtain an explicit expression for the increase in MSE of the new GREG estimator relative to the MSE of the BLUP estimator. In Section 4, we provide empirical results on the relative efficiency of the estimators when the model parameters are estimated. We used a model-based set-up as

well as a real finite population for the simulation study. Finally, some concluding remarks are given in Section 5.

## 2. BLUP and GREG estimators

The two-level model (2) is a special case of a general linear mixed model with block diagonal covariance structure. Therefore, assuming known model parameters, we may calculate the BLUP estimator of  $\mu_i$  as

$$\tilde{\mu}_i^B = \bar{X}_i'(Z_i\beta + \tilde{v}_i), \quad (3)$$

where

$$\tilde{v}_i = \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta),$$

and the superscript B on  $\tilde{\mu}_i$  stands for BLUP estimator (Rao 2003, page 107). Similarly, the BLUP estimator of a non-sampled observation  $j$  in  $i^{\text{th}}$  area can be written as

$$\tilde{y}_{ij} = x_{ij}'(Z_i\beta + \tilde{v}_i). \quad (4)$$

On the other hand, a model-assisted GREG estimator of  $\mu_i$  (or  $\bar{Y}_i$ ) is given by

$$\tilde{\mu}_i^G = \frac{1}{N_i} \left[ \sum_{j=1}^{N_i} \tilde{y}_{ij} + \sum_{j=1}^{n_i} w_{ij} (y_{ij} - \tilde{y}_{ij}) \right], \quad i = 1, \dots, m \quad (5)$$

where  $\tilde{y}_{ij}$  is the predictor of  $y_{ij}$  under the assumed model, and  $w_{ij}$  is the survey weight which equals  $N_i/n_i$  in the case of simple random sampling (SRS) within areas. We focus on SRS within areas in this paper.

Using (5) with  $\tilde{y}_{ij} = x_{ij}' Z_i \beta$  as the predictor of  $y_{ij}$  under the model (1) with fixed  $\beta_i = Z_i \beta$ , we can write the customary GREG estimator as

$$\tilde{\mu}_i^G = \bar{y}_i + (\bar{X}_i - \bar{x}_i)' Z_i \beta, \quad (6)$$

where the superscript G on  $\tilde{\mu}_i$  denotes GREG (Särndal, Swensson and Wretman 1992, page 225),  $\bar{y}_i$  is sample mean of  $y_{ij}$  in the  $i^{\text{th}}$  area, and  $\bar{x}_i$  is the sample mean of  $x_{ij}$  in the  $i^{\text{th}}$  area, respectively. Using the predictor (4) based on the two-level model (1) in (5), we get a new GREG estimator of  $\mu_i$  (or  $\bar{Y}_i$ ) as

$$\begin{aligned} \tilde{\mu}_i^{\text{LV}} &= [\bar{X}_i'(Z_i\beta + \tilde{v}_i) + (\bar{y}_i - \bar{x}_i)'(Z_i\beta + \tilde{v}_i)] \\ &= \bar{y}_i + (\bar{X}_i - \bar{x}_i)'(Z_i\beta + \tilde{v}_i), \end{aligned} \quad (7)$$

where the superscript LV on  $\tilde{\mu}_i$  denotes that it was first proposed by Lehtonen and Veijanen (1999). The estimators  $\tilde{\mu}_i^B$ ,  $\tilde{\mu}_i^G$  and  $\tilde{\mu}_i^{\text{LV}}$  are linear in the  $y_{ij}$ 's and unbiased under the two-level model (1). In practice, we replace the parameters  $\beta$ ,  $\Sigma_v$  and  $\sigma_e^2$  in (3), (6) and (7) by suitable estimators. The resulting estimators are denoted by  $\hat{\mu}_i^B$ ,  $\hat{\mu}_i^G$  and  $\hat{\mu}_i^{\text{LV}}$  respectively, where  $\hat{\mu}_i^B$  is the empirical BLUP (EBLUP) estimator. Under normality assumption,  $\hat{\mu}_i^B$  is the empirical best (EB) estimator. The EBLUP estimator of  $\bar{Y}_i$  is given in Section 4.2.2. Note that  $\hat{\mu}_i^G$  and  $\hat{\mu}_i^{\text{LV}}$  are valid as estimators of  $\bar{Y}_i$ .

## 3. Mean squared error

The mean squared error (MSE) of the customary GREG estimator  $\tilde{\mu}_i^G$  under the two-level model can be written as

$$\begin{aligned} \text{MSE}(\tilde{\mu}_i^G) &= E(\tilde{\mu}_i^G - \mu_i)^2 \\ &= E[\bar{y}_i + (\bar{X}_i - \bar{x}_i)' Z_i \beta - \bar{X}_i'(Z_i\beta + \tilde{v}_i)]^2 \\ &= E[(\bar{x}_i - \bar{X}_i)' \tilde{v}_i + \bar{e}_i]^2 \\ &= (\bar{x}_i - \bar{X}_i)' \Sigma_v (\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i}, \end{aligned}$$

as stated in Theorem 1.

Theorem 1. *The MSE of the GREG estimator (6) is given by*

$$\text{MSE}(\tilde{\mu}_i^G) = (\bar{x}_i - \bar{X}_i)' \Sigma_v (\bar{x}_i - \bar{X}_i) + \frac{\sigma_e^2}{n_i}. \quad (8)$$

Further, we may write the MSE of the BLUP estimator  $\tilde{\mu}_i^B$  as follows:

$$\begin{aligned} \text{MSE}(\tilde{\mu}_i^B) &= E(\tilde{\mu}_i^B - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{v}_i - \tilde{v}_i)]^2 \\ &= \bar{X}_i'(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v) \bar{X}_i, \end{aligned}$$

as stated in Theorem 2.

Theorem 2. *The MSE of the BLUP estimator (3) is given by*

$$\text{MSE}(\tilde{\mu}_i^B) = \bar{X}_i'(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v) \bar{X}_i. \quad (9)$$

Theorem 3 gives the MSE of the new GREG estimator  $\tilde{\mu}_i^{\text{LV}}$ .

Theorem 3. *The MSE of the new GREG estimator (7) is given by*

$$\begin{aligned} \text{MSE}(\tilde{\mu}_i^{\text{LV}}) &= \text{MSE}(\tilde{\mu}_i^B) \\ &+ \left\{ \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \right\}. \end{aligned} \quad (10)$$

Proof of Theorem 3 is given in the Appendix.

By definition, we have  $\text{MSE}(\tilde{\mu}_i^B) \leq \text{MSE}(\tilde{\mu}_i^{\text{LV}})$  and (10) gives an explicit expression for the increase in MSE of  $\tilde{\mu}_i^{\text{LV}}$  relative to the MSE of the BLUP estimator  $\tilde{\mu}_i^B$ .

## 4. Empirical results

### 4.1 Empirical comparison of MSE values

In order to study the efficiency of the new GREG estimator, we used data from Moura and Holt (1999) based on 38,740 households in the enumeration districts (small areas) in one county in Brazil. The income of household's head was treated as the response variable  $y$ . Two unit level independent variables were identified as the educational attainment of household's head (ordinal scale of 0-5) and the number of rooms in the household (1-11+). The following two-level model was assumed for this data:



$$y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + e_{ij},$$
$$j = 1, \dots, N_i; i = 1, \dots, m \tag{11}$$

with

$$\beta_{i0} = \beta_0 + v_{i0}; \beta_{i1} = \beta_1 + v_{i1}; \beta_{i2} = \beta_2 + v_{i2}, \tag{12}$$

where

$$v = (v_{i0}, v_{i1}, v_{i2})' \sim N_3(0, \Sigma_v), e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$$

and  $x_1$  and  $x_2$  respectively represent the number of rooms and the educational attainment of household's head (centered about their respective population means). Note that the model (12) for the random  $\beta_i$ -coefficients does not contain area level covariates  $Z$ .

Moura and Holt (1999) also studied another model with an area level covariate  $Z$  for modeling  $\beta_i$ 's in (12). For this data, average number of cars per household in each area was used as a covariate  $z$  for modeling the random coefficients  $\beta_{i1}$  and  $\beta_{i2}$  corresponding to the variables  $x_1$  and  $x_2$ , but not for the random intercept term,  $\beta_{i0}$ . The two-level model (11) with the area level covariate  $z$  is given by

$$y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij},$$
$$j = 1, \dots, N_i; i = 1, \dots, m \tag{13}$$

with

$$\beta_{i0} = \beta_0 + v_{i0};$$
$$\beta_{i1} = \beta_1 + \alpha_1 z_i + v_{i1}; \beta_{i2} = \beta_2 + \alpha_2 z_i + v_{i2}. \tag{14}$$

Moura and Holt (1999) fitted models (11)-(12) and (13)-(14) to the full data set mentioned above. We summarize their results in Table 1.

**Table 1**  
Parameter estimates based on Moura and Holt's (1999) data set, where  $\sigma_0^2$ ,  $\sigma_1^2$  and  $\sigma_2^2$  are the diagonal elements and  $\sigma_{01}$ ,  $\sigma_{02}$  and  $\sigma_{12}$  are the off-diagonal elements of the covariance matrix  $\Sigma_v$

Parameter	Diagonal Covariance: Model (14) with $z$	Diagonal Covariance: Model (12) without $z$	General Covariance: Model (12) without $z$
$\beta_0$	8.442	8.688	8.456
$\beta_1$	0.451	1.321	1.223
$\beta_2$	0.744	2.636	2.596
$\alpha_1$	3.779	-	-
$\alpha_2$	1.659	-	-
$\sigma_0^2$	0.745	0.637	1.385
$\sigma_1^2$	0.237	0.471	0.234
$\sigma_2^2$	0.700	1.472	0.926
$\sigma_{01}$	-	-	0.354
$\sigma_{02}$	-	-	0.492
$\sigma_{12}$	-	-	0.333
$\sigma_e^2$	44.00	44.01	47.74

The area means of  $x_1$  and  $x_2$  were calculated from the whole data and treated as the population means  $\bar{X}_{1i}$  and

$\bar{X}_{2i}$ . A random subsample of 10% of the records was selected from each small area. The overall sample size is  $n = 3,876$  and the number of small areas is  $m = 140$ . Using the sample values of  $x_1$ ,  $x_2$  and  $z$  and the population means  $\bar{X}_{1i}$  and  $\bar{X}_{2i}$ , we computed  $MSE(\hat{\mu}_i^G)$ ,  $MSE(\hat{\mu}_i^B)$  and  $MSE(\hat{\mu}_i^{LV})$  using (8), (9) and (10) respectively, treating the estimates of regression parameters,  $\Sigma_v$  and  $\sigma_e^2$  in Table 1 for the full data as true values. We then calculated the average MSE values over the areas:

$$\overline{MSE}_G = \frac{1}{m} \sum_{i=1}^m MSE(\hat{\mu}_i^G),$$
$$\overline{MSE}_B = \frac{1}{m} \sum_{i=1}^m MSE(\hat{\mu}_i^B)$$

and

$$\overline{MSE}_{LV} = \frac{1}{m} \sum_{i=1}^m MSE(\hat{\mu}_i^{LV}).$$

We define the relative efficiency of  $\hat{\mu}^B$  over  $\hat{\mu}^G$  as  $EFF_B$  and the relative efficiency of  $\hat{\mu}^{LV}$  over  $\hat{\mu}^G$  as  $EFF_{LV}$ , where

$$EFF_B = \frac{\overline{MSE}_G}{\overline{MSE}_B}; \quad EFF_{LV} = \frac{\overline{MSE}_G}{\overline{MSE}_{LV}}.$$

We summarize the results in Tables 2 and 3. Tables 2 and 3 reveal that the new GREG estimator is slightly more efficient than the usual GREG estimator in terms of average MSE:  $EFF_{LV} \leq 112\%$ . However, the new GREG estimator is substantially less efficient than the BLUP estimator, under the assumed two-level model. For example, for the model with  $z$  and diagonal covariance matrix (Table 2),  $EFF_B = 292\%$  compared to  $EFF_{LV} = 106\%$ , and  $MSE_B = 0.62$  compared to  $\overline{MSE}_{LV} = 1.72$ .

**Table 2**  
Comparison of small area estimators: relative efficiency (EFF) and average MSE (MSE) for the case of diagonal covariance matrix based on Moura and Holt's (1999) data set

Quality Measure	Model without $z$			Model with $z$		
	GREG	New GREG	BLUP	GREG	New GREG	BLUP
EFF	100%	112%	306%	100%	106%	292%
MSE	1.92	1.71	0.62	1.83	1.72	0.62

**Table 3**  
Comparison of small area estimators: relative efficiency (EFF) and average MSE (MSE) for the case of a general covariance matrix based on Moura and Holt's (1999) data set

Quality Measure	Model without $z$		
	GREG	New GREG	BLUP
EFF	100%	108%	253%
MSE	2.02	1.87	0.80

4.2 Simulation study

4.2.1 Simulation study under a model-based framework

In order to investigate the efficiency of the new GREG estimator with estimated model parameters, a small

simulation study based on the two-level models (11)-(12) and (13)-(14) was undertaken. We only considered a diagonal covariance structure  $\Sigma_v$  with diagonal elements  $\sigma_{0s}^2$ ,  $\sigma_1^2$  and  $\sigma_2^2$ . We again used the data from Moura and Holt (1999). The estimates of  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $\alpha_1$ ,  $\alpha_2$ ,  $\sigma_0^2$ ,  $\sigma_1^2$ ,  $\sigma_2^2$  and  $\sigma_e^2$  reported in Table 1 are treated as true values.

In our simulation study, we took  $(x_{1ij}, x_{2ij}, z_i)$  from Moura and Holt (1999) and then generated  $y_{ij}$  based on the models (11)-(12) and (13)-(14). By using the generated samples  $(y_{ij}^{(b)}, x_{1ij}, x_{2ij}, z_i)$ ,  $b = 1, \dots, B = 1,000$ , we calculated  $\hat{\beta}^{(b)}$  by generalized least squares for the new GREG method as well as for the BLUP method. For the GREG method we used ordinary least squares to estimate  $\beta$  as  $\hat{\beta}_{ols}^{(b)}$ . In addition,  $\hat{\Sigma}_v^{(b)}$  and  $\hat{\sigma}_e^{2(b)}$  were computed based on the restricted maximum likelihood (REML) method. For each generated sample, we calculated

$$\mu_i^{(b)} = \bar{X}'(Z_i\beta + v_i^{(b)}), \quad i = 1, \dots, m; \quad b = 1, \dots, B.$$

We computed the new GREG estimator of  $\mu_i^{(b)}$  as  $\hat{\mu}_i^{LV(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})$ , the GREG estimator of  $\mu_i^{(b)}$  as  $\hat{\mu}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)'Z_i\hat{\beta}_{ols}^{(b)}$  and the empirical BLUP (EBLUP) estimator of  $\mu_i^{(b)}$  as  $\hat{\mu}_i^{B(b)} = \bar{X}_i'(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})$ , where  $\hat{v}_i^{(b)} = \hat{\Sigma}_v^{(b)}X_i'\hat{V}_i^{-1(b)}(y_i^{(b)} - X_iZ_i\hat{\beta}^{(b)})$ .

We then computed the average mean squared error ( $\overline{MSE}_1$ ) and average absolute relative error ( $\overline{ARE}_1$ )

$$\overline{MSE}_1 = \frac{1}{m} \sum_i MSE_{1i} \quad \text{where} \quad MSE_{1i} = B^{-1} \sum_{b=1}^B (\hat{\mu}_i^{(b)} - \mu_i^{(b)})^2,$$

$$\overline{ARE}_1 = \frac{1}{m} \sum_i ARE_{1i} \quad \text{where} \quad ARE_{1i} = B^{-1} \sum_{b=1}^B |\hat{\mu}_i^{(b)} - \mu_i^{(b)}| / \mu_i^{(b)},$$

where  $\hat{\mu}_i^{(b)}$  denotes  $\hat{\mu}_i^{LV(b)}$ ,  $\hat{\mu}_i^{G(b)}$  or  $\hat{\mu}_i^{B(b)}$ . We report the results in Table 4. Both models with area level covariate  $z$  and without  $z$  have slightly smaller values of  $\overline{MSE}_1$  and  $\overline{ARE}_1$  for the new GREG estimator relative to the GREG estimator. However,  $\overline{MSE}_1$  and  $\overline{ARE}_1$  are significantly smaller for the EBLUP estimator due to borrowing strength from related areas. Moreover, comparing Tables 2 and 4, we can see that the values of  $\overline{MSE}_1$  in Table 4 are slightly larger than the corresponding values in Table 2 due to estimating model parameters.

**Table 4**  
Comparison of small area estimators: average MSE ( $\overline{MSE}_1$ ) and average absolute relative error ( $\overline{ARE}_1$ ) under a model-based framework

Quality Measure	Model without $z$			Model with $z$		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{MSE}_1$	1.93	1.73	0.67	1.84	1.73	0.73
$\overline{ARE}_1$	0.14	0.13	0.08	0.13	0.12	0.08

## 4.2.2 Simulation study under a finite population framework

To study the performance of the estimators under a finite population framework, we created a synthetic finite population from the Brazilian data consisting of  $n = 3,876$  sample values  $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$ . By duplicating the sample values  $(y_{ij}, x_{1ij}, x_{2ij}, z_i)$  five times, we treated the new  $(y, x_1, x_2, z)$ -data of size 19,380 as our real population.

We generated 500 independent samples ( $B = 500$ ), each of size  $n = 700$  and  $n = 1,400$ , by taking simple random samples of size  $n_i = 5$  and  $n_i = 10$  in each area  $i = 1, \dots, 140$ . As before, for each sample we calculated  $\hat{\beta}^{(b)}$  for the new GREG and the BLUP methods and  $\hat{\beta}_{ols}^{(b)}$  for the GREG method. In addition,  $\hat{\Sigma}_v^{(b)}$  and  $\hat{\sigma}_e^{2(b)}$  were calculated based on the REML method. We also computed the population mean of  $y_{ij}$  for each area  $i$  as

$$\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i, \quad i = 1, \dots, 140,$$

where  $N_i$  is the population size in  $i^{\text{th}}$  area. Further, for each sample  $b = 1, \dots, B$ , we calculated the new GREG estimate of the  $i^{\text{th}}$  area mean as  $\hat{Y}_i^{LV(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})$ , the GREG estimate as  $\hat{Y}_i^{G(b)} = \bar{y}_i^{(b)} + (\bar{X}_i - \bar{x}_i)'Z_i\hat{\beta}_{ols}^{(b)}$  and the EBLUP estimate as  $\hat{Y}_i^{B(b)} = f_i\bar{y}_i + (1 - f_i)[\bar{X}_i'(Z_i\hat{\beta}^{(b)} + \hat{v}_i^{(b)})]$ , where

$$f_i = n_i / N_i, \quad \bar{X}_i^* = \frac{N_i \bar{X}_i - n_i \bar{x}_i}{N_i - n_i},$$

and  $\hat{v}_i^{(b)} = \hat{\Sigma}_v^{(b)}X_i'\hat{V}_i^{-1(b)}(y_i^{(b)} - X_iZ_i\hat{\beta}^{(b)})$ .

The EBLUP estimator accounts for the finite population corrections  $f_i$ .

We computed the average mean squared error ( $\overline{MSE}_2$ ) and average absolute relative error ( $\overline{ARE}_2$ ) as

$$\overline{MSE}_2 = \frac{1}{m} \sum_i MSE_{2i}, \quad \overline{ARE}_2 = \frac{1}{m} \sum_i ARE_{2i},$$

where

$$MSE_{2i} = \frac{1}{B} \sum_{b=1}^B (\hat{Y}_i^{(b)} - \bar{Y}_i)^2, \quad ARE_{2i} = \frac{1}{B} \sum_{b=1}^B |\hat{Y}_i^{(b)} - \bar{Y}_i| / \bar{Y}_i$$

and  $\hat{Y}_i^{(b)}$  denotes  $\hat{Y}_i^{LV(b)}$ ,  $\hat{Y}_i^{G(b)}$  or  $\hat{Y}_i^{B(b)}$ . We report the results in Tables 5 and 6 for  $n_i = 5$  and  $n_i = 10$  respectively. Both models with area level covariate  $z$  and without  $z$  are considered.

**Table 5**  
Comparison of small area estimators: average MSE ( $\overline{MSE}_2$ ) and average absolute relative error ( $\overline{ARE}_2$ ) under a finite population framework ( $n_i = 5$ )

Quality Measure	Model without $z$			Model with $z$		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{MSE}_2$	11.03	10.02	6.50	10.76	10.06	7.06
$\overline{ARE}_2$	0.27	0.24	0.18	0.25	0.23	0.22



Table 5 shows that for  $n_i = 5$  the new GREG estimator exhibits slightly better performance relative to the GREG estimator in the sense of smaller  $\overline{\text{MSE}}_2$  and  $\overline{\text{ARE}}_2$ . On the other hand, Table 6 reveals that with  $n_i = 10$  the GREG estimator has slightly better performance than the new GREG estimator in terms of  $\text{MSE}_2$  but not  $\overline{\text{ARE}}_2$ . However, the EBLUP estimator gives substantially smaller  $\overline{\text{MSE}}_2$  and  $\overline{\text{ARE}}_2$  than the GREG and the new GREG in both cases due to borrowing strength from related small areas. For example, for the model without  $z$  and  $n_i = 5$ ,  $\overline{\text{MSE}}_2 = 10.02$ ,  $11.03$  and  $6.50$  for the new GREG, the GREG and the EBLUP, respectively.

**Table 6**  
Comparison of small area estimators: average MSE ( $\overline{\text{MSE}}_2$ ) and average absolute relative error ( $\overline{\text{ARE}}_2$ ) under finite population framework ( $n_i = 10$ )

Quality Measure	Model without $z$			Model with $z$		
	GREG	New GREG	EBLUP	GREG	New GREG	EBLUP
$\overline{\text{MSE}}_2$	6.53	6.77	4.73	6.75	6.96	5.24
$\overline{\text{ARE}}_2$	0.20	0.18	0.15	0.19	0.18	0.19

## 5. Summary

In this paper, we derived the model mean squared error (MSE) of a two-level model-assisted new GREG estimator of a small area mean, proposed by Lehtonen and Veijanen (1999). In addition, we used a data set of Moura and Holt (1999) to demonstrate empirically that the BLUP estimator is substantially more efficient than the new GREG estimator in terms of model MSE, while the new GREG is only slightly more efficient than the customary GREG based on the regression model  $y_i = X_i Z_i \beta + e_i$ ,  $i = 1, \dots, m$ . Moreover, using a simulation study under a model-based framework, we have shown that the new GREG estimator has consistently better performance relative to the GREG estimator in terms of average MSE,  $\overline{\text{MSE}}$ , and average absolute relative error,  $\overline{\text{ARE}}$ . However, due to borrowing strength from related small areas, EBLUP estimator exhibits significantly better performance relative to the GREG and the new GREG estimators. In addition, we conducted a simulation study under a finite population framework and showed that the EBLUP estimator outperforms the new GREG and the GREG estimators in terms of  $\overline{\text{MSE}}$  and  $\overline{\text{ARE}}$ .

## Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. It is based on a chapter in M.Torabi's Ph.D. thesis written under

the supervision of J.N.K. Rao. The authors thank the referees and an associate editor for their helpful comments on the original version of this paper.

## Appendix

### Derivation of $\text{MSE}(\hat{\mu}_i^{\text{LV}})$ :

$$\begin{aligned} \text{MSE}(\hat{\mu}_i^{\text{LV}}) &= E(\hat{\mu}_i^{\text{LV}} - \mu_i)^2 \\ &= E[\bar{X}_i'(\tilde{y}_i - v_i)]^2 + E[\bar{y}_i - \bar{x}_i'(Z_i \beta + \tilde{v}_i)]^2 \\ &\quad + 2E[(\bar{y}_i - \bar{x}_i'(Z_i \beta + \tilde{v}_i))\bar{X}_i'(\tilde{y}_i - v_i)], \end{aligned} \quad (\text{A.1})$$

where the first term on the right hand side of (A.1) is the MSE of the BLUP estimator under the two-level model, given by (9). Moreover, we may write

$$\begin{aligned} E[\bar{y}_i - \bar{x}_i'(Z_i \beta + \tilde{v}_i)]^2 &= E[\bar{y}_i - \bar{x}_i' Z_i \beta]^2 \\ &\quad + E(\bar{x}_i' \tilde{v}_i)^2 - 2E[(\bar{y}_i - \bar{x}_i' Z_i \beta)(\bar{x}_i' \tilde{v}_i)], \end{aligned} \quad (\text{A.2})$$

where

$$E[\bar{y}_i - \bar{x}_i' Z_i \beta]^2 = \text{Var}(\bar{y}_i) = \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i},$$

and

$$\begin{aligned} E(\bar{x}_i' \tilde{v}_i)^2 &= \text{Var}(\bar{x}_i' \tilde{v}_i) + [E(\bar{x}_i' \tilde{v}_i)]^2 \\ &= \text{Var}[\bar{x}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)] \\ &\quad + [E(\bar{x}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta))]^2 \\ &= \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i. \end{aligned}$$

In addition,

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i' Z_i \beta)(\bar{x}_i' \tilde{v}_i)] &= E[\bar{y}_i \bar{x}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)] \\ &\quad - E[\bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)], \end{aligned}$$

where the second term is zero. Therefore, we may write

$$\begin{aligned} E[(\bar{y}_i - \bar{x}_i' Z_i \beta)(\bar{x}_i' \tilde{v}_i)] &= E[\bar{y}_i \bar{x}_i' \Sigma_v X_i' V_i^{-1} y_i] \\ &\quad - \bar{x}_i' Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta, \end{aligned}$$

where the first term can be written as

$$\begin{aligned} E\left[\frac{1}{n_i} y_i \bar{x}_i' \Sigma_v X_i' V_i^{-1} y_i\right] &= \\ \frac{1}{n_i} (X_i \Sigma_v \bar{x}_i + X_i Z_i \beta \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta), \end{aligned}$$

using the following Lemma:

LEMMA 1 (Searle 1971). *If  $y$  is a  $n \times 1$  vector with mean  $\mu$  and variance-covariance matrix  $\Sigma$  and  $b$  is a  $n \times 1$  vector, then  $E(yb'y) = \Sigma b + \mu b'\mu$ .*

Hence,

$$\begin{aligned} E[(\bar{y}_i - \bar{x}'_i Z_i \beta)(\bar{x}'_i \tilde{v}_i)] \\ = \bar{x}'_i \Sigma_v \bar{x}_i + \bar{x}'_i Z_i \beta \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ - \bar{x}'_i Z_i \beta \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \\ = \bar{x}'_i \Sigma_v \bar{x}_i. \end{aligned}$$

Then we may write (A.2) as follows:

$$\begin{aligned} E[\bar{y}_i - \bar{x}'_i (Z_i \beta + \tilde{v}_i)]^2 &= \bar{x}'_i \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \\ &+ \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - 2 \bar{x}'_i \Sigma_v \bar{x}_i \\ &= \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}'_i \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i}. \quad (A.3) \end{aligned}$$

Finally, we need to find the cross-product term of (A.1). We have

$$\begin{aligned} E[(\bar{y}_i - \bar{x}'_i (Z_i \beta + \tilde{v}_i)) \bar{X}'_i (\tilde{v}_i - v_i)] \\ = E[\bar{y}_i \bar{X}'_i (\tilde{v}_i - v_i)] \\ - E[\bar{x}'_i (Z_i \beta + \tilde{v}_i) \bar{X}'_i (\tilde{v}_i - v_i)], \quad (A.4) \end{aligned}$$

where the first term on the right side of (A.4) may be written as

$$\begin{aligned} E[\bar{y}_i \bar{X}'_i (\tilde{v}_i - v_i)] \\ = E[\bar{y}_i \bar{X}'_i (\Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta) - v_i)] \\ = E[\bar{y}_i \bar{X}'_i \Sigma_v X_i' V_i^{-1} y_i] \\ - E[y_i \bar{X}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta] \\ - E[\bar{y}_i \bar{X}'_i v_i]. \quad (A.5) \end{aligned}$$

The first term of (A.5) is

$$\begin{aligned} E[\bar{y}_i \bar{X}'_i \Sigma_v X_i' V_i^{-1} y_i] \\ = E\left[\frac{1}{n_i} y_i \bar{X}'_i \Sigma_v X_i' V_i^{-1} y_i\right] \\ = \frac{1}{n_i} X_i (\Sigma_v \bar{X}_i + Z_i \beta \bar{X}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta), \quad (A.6) \end{aligned}$$

the second term of (A.5) can be written as

$$\begin{aligned} E[\bar{y}_i \bar{X}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta] \\ = \bar{x}'_i Z_i \beta \bar{X}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta, \quad (A.7) \end{aligned}$$

and the third term can be expressed as

$$\begin{aligned} E[\bar{y}_i \bar{X}'_i v_i] &= E[(\bar{x}'_i Z_i \beta + \bar{x}'_i v_i + e_i) \bar{X}'_i v_i] \\ &= E[\bar{x}'_i v_i \bar{X}'_i v_i] = \bar{x}'_i \Sigma_v \bar{X}_i. \quad (A.8) \end{aligned}$$

Therefore, substituting (A.6), (A.7) and (A.8) in (A.5), we have

$$E[\bar{y}_i \bar{X}'_i (\tilde{v}_i - v_i)] = 0. \quad (A.9)$$

We now turn to the second term of (A.4). We have

$$\begin{aligned} E[\bar{x}'_i (Z_i \beta + \tilde{v}_i) \bar{X}'_i (\tilde{v}_i - v_i)] \\ = E[\bar{x}'_i Z_i \beta \bar{X}'_i \tilde{v}_i] + E[\bar{x}'_i \tilde{v}_i \bar{X}'_i \tilde{v}_i] \\ - E[\bar{x}'_i Z_i \beta \bar{X}'_i v_i] - E[\bar{x}'_i \tilde{v}_i \bar{X}'_i v_i]. \quad (A.10) \end{aligned}$$

Then we obtain the following expression for the four terms on the right side of (A.10):

$$\begin{aligned} E[\bar{x}'_i Z_i \beta \bar{X}'_i \tilde{v}_i] \\ = E[\bar{x}'_i Z_i \beta \bar{X}'_i \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)] = 0, \quad (A.11) \end{aligned}$$

$$\begin{aligned} E[\bar{x}'_i \tilde{v}_i \bar{X}'_i \tilde{v}_i] &= \bar{x}'_i \Sigma_v \bar{X}_i + \bar{x}'_i E(\tilde{v}_i) \bar{X}'_i E(\tilde{v}_i) \\ &= \bar{x}'_i \text{Var}[\Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)] \bar{X}_i \\ &+ \bar{x}'_i E(\Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)) \\ &\times \bar{X}'_i E(\Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta)) \\ &= \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i, \quad (A.12) \end{aligned}$$

$$E[\bar{x}'_i Z_i \beta \bar{X}'_i v_i] = 0 \quad (A.13)$$

and

$$\begin{aligned} E[\bar{x}'_i \tilde{v}_i \bar{X}'_i v_i] \\ = E[\bar{x}'_i \Sigma_v X_i' V_i^{-1} (y_i - X_i Z_i \beta) \bar{X}'_i v_i] \\ = E[\bar{x}'_i \Sigma_v X_i' V_i^{-1} y_i \bar{X}'_i v_i] \\ - E[\bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i Z_i \beta \bar{X}'_i v_i] \\ = E[\bar{x}'_i \Sigma_v X_i' V_i^{-1} (X_i Z_i \beta + X_i v_i + e_i) \bar{X}'_i v_i] \\ = E[\bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i v_i \bar{X}'_i v_i] \\ = \bar{x}'_i \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i. \quad (A.14) \end{aligned}$$

Therefore, substituting (A.11)-(A.14) in (A.10), we get

$$E[\bar{x}'_i (Z_i \beta + \tilde{v}_i) \bar{X}'_i (\tilde{v}_i - v_i)] = 0. \quad (A.15)$$

Hence, it follows from (A.4), (A.9) and (A.15) that

$$E[(y_i - \bar{x}'_i (Z_i \beta + \tilde{v}_i)) \bar{X}'_i (\tilde{v}_i - v_i)] = 0. \quad (A.16)$$

It now follows from (A.1), (A.3) and (A.16) that



$$\text{MSE}(\tilde{\mu}_i^{\text{LV}}) = \text{MSE}(\tilde{\mu}_i^{\beta}) + \left\{ \bar{x}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{x}_i - \bar{x}_i' \Sigma_v \bar{x}_i + \frac{\sigma_e^2}{n_i} \right\},$$

as stated in Theorem 3.

## References

Lehtonen, R., and Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *IASS Satellite Conference on Small Area Estimation*, Riga: Latvian Council of Science, 121- 128.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29, 33-44.

Moura, F.A.S., and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology*, 25, 73-80.

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken: New York: John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Searle, S.R. (1971). *Linear Models*. New York: John Wiley & Sons, Inc.





# An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada

Yong You<sup>1</sup>

## Abstract

The Canadian Labour Force Survey (LFS) produces monthly estimates of the unemployment rate at national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Urban Centers (UCs). However, for some sub-provincial areas, the direct estimates are not reliable since the sample size in some areas is quite small. The small area estimation in LFS concerns estimation of unemployment rates for local sub-provincial areas such as CMA/UCs using small area models. In this paper, we will discuss various models including the Fay-Herriot model and cross-sectional and time series models. In particular, an integrated non-linear mixed effects model will be proposed under the hierarchical Bayes (HB) framework for the LFS unemployment rate estimation. Monthly Employment Insurance (EI) beneficiary data at the CMA/UC level are used as auxiliary covariates in the model. A HB approach with the Gibbs sampling method is used to obtain the estimates of posterior means and posterior variances of the CMA/UC level unemployment rates. The proposed HB model leads to reliable model-based estimates in terms of CV reduction. Model fit analysis and comparison of the model-based estimates with the direct estimates are presented in the paper.

**Key Words:** Design effect; Hierarchical Bayes; Log-linear mixed effects model; Model checking; Sampling variance; Small area.

## 1. Introduction

The unemployment rate is generally viewed as a key indicator of economic performance. In Canada, the unemployment rate estimates are produced monthly by the Labour Force Survey (LFS) of Statistics Canada. The LFS is a monthly survey of 53,000 households selected using a stratified, multistage design. Each month, one-sixth of the sample is replaced. Thus five-sixths of the sample is common between two consecutive months. This sample overlap induces correlations which can be exploited to produce better estimates by alternative methods such as model-based methods to borrow strength over time; more details will be discussed in Section 2. For a detailed description of the LFS design, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998). The LFS releases monthly unemployment rate estimates for large areas such as the nation and provinces as well as local areas (small areas) such as Census Metropolitan Areas (CMAs, *i.e.*, cities with population more than 100,000) and other Urban Centres (UCs) across Canada. Although national and provincial estimates get the most media attention, sub-provincial estimates of the unemployment rate are also very important. They are used by the Employment Insurance (EI) program to determine the rules used to administer the program. In addition, the unemployment rates for CMAs and UCs receive close scrutiny at local levels. However, many local areas do not have large enough samples to produce adequate direct estimates, since the LFS is designed to produce adequate or reliable estimates at the national level and

provincial level. The estimated coefficient of variation (CV) level for the nation is about 2% and 4% to 7% for provinces. However, the CVs for CMAs and UCs range from about 7% to 50%. Some UCs have CVs even larger than 50%. The direct LFS estimates for some local areas are not reliable with very large CVs due to the small sample sizes for those areas. Therefore, alternative estimators, in particular, model-based estimators, are considered to improve the direct LFS estimates for small areas. The objective in this paper is to obtain a reliable model-based estimator that is an improvement over the direct LFS estimator in terms of small and stable CVs.

In general, direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains such as the nation and provinces. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide reliable direct estimates for specific small domains. In making estimates for small areas, it is necessary to borrow strength from related areas to form indirect estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as census counts and administrative records. It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data; see Rao (2003). The model-based estimators are indirect estimators in the sense that these estimators are obtained by using small

1. Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: yong.you@statcan.ca.

area models, direct estimates and related auxiliary variables. The model-based estimators are obtained to improve the direct design-based estimators in terms of precision and reliability, that is, smaller CVs. Traditional small area estimators borrow strength either from similar small areas or from the same area across time, but not both. In recent years, several approaches to borrowing strength simultaneously across both space and time have been developed. Estimators based on the approach developed by Rao and Yu (1994), Ghosh, Nangia and Kim (1996), Datta, Lahiri, Maiti and Lu (1999), Datta, Lahiri and Maiti (2002) and You, Rao and Gambino (2000, 2003), successfully exploit the two dimensions simultaneously to produce improved estimates with desirable properties for small areas. In particular, You *et al.* (2000, 2003) studied the model-based estimation of unemployment rates for local sub-provincial areas such as CMAs and Census Agglomerations (CAs) across Canada. They obtained efficient model-based estimators and adequate model fit for the LFS unemployment rate estimation. However, the model proposed by You *et al.* (2000, 2003) has some limitations. In this paper, we discuss these limitations and propose a new integrated model for the LFS unemployment rate estimation under hierarchical Bayes (HB) framework. The idea is to model the parameters of interest and the sampling variances together as suggested in You *et al.* (2003) and You and Chapman (2006). We will apply the proposed model to the 2005 LFS data and obtain the model-based unemployment rate estimates. Comparison of the HB estimates with the direct LFS estimates and model fit analysis will also be provided.

This paper is organized as follows. In Section 2, we present and discuss various small area models proposed in the literature for the unemployment rate estimation. In Section 3, we discuss the problem of smoothing and modeling the sampling covariance matrix. In Section 4, an integrated non-linear mixed effects model is proposed in a hierarchical Bayes framework, and the use of Gibbs sampling to generate samples from the joint posterior distribution is described. In Sections 5, we apply the proposed model to LFS data and obtain the HB estimates for small area unemployment rates. Model analysis and evaluation are also provided. And finally in Section 6 we offer some concluding remarks and future work directions.

## 2. Small area models

### 2.1 Cross-sectional model

Cross-sectional or area level models are used to produce reliable model-based estimates by combining area level auxiliary information and direct area level estimates. A basic area level model is the well-known Fay-Herriot model

(Fay and Herriot 1979). This model has two components: (1) a sampling model for the direct survey estimates, and (2) a linking model that relates the small area parameters to area level auxiliary variables through a linear regression model. For the LFS monthly unemployment rate estimation, let  $\theta_{it}$  denote the true unemployment rate for the  $i^{\text{th}}$  CMA/UC at a particular time (month)  $t$ , where  $i=1, \dots, m$ , where  $m$  is the number of CMA/UCs, and let  $y_{it}$  denote the direct LFS estimate of  $\theta_{it}$ . Then the sampling model for  $y_{it}$  can be expressed as

$$y_{it} = \theta_{it} + e_{it}, \quad i=1, \dots, m, \quad (1)$$

where  $e_{it}$  is the sampling error associated with the direct estimator  $y_{it}$ . The sampling error is assumed to be normally distributed as  $e_{it} \sim N(0, \sigma_{it}^2)$  where  $\sigma_{it}^2$  is the sampling variance. The linking model for the true unemployment rate  $\theta_{it}$  may be written as

$$\theta_{it} = x'_{it}\beta + v_i, \quad i=1, \dots, m, \quad (2)$$

where  $x_{it}$  is the auxiliary variable and  $v_i$  is area-specific random effect. For each time point (each month), we can use the Fay-Herriot model for the monthly direct estimates. The Fay-Herriot model combines cross-sectional information but does not borrow strength over the past time periods.

### 2.2 Cross-sectional and time series model

Because of the LFS sample design and rotation pattern, there is substantial sample overlap over six month time periods within each area. As a result, for a particular area  $i$ , the correlation between the sampling errors  $e_{it}$  and  $e_{is}$  ( $t \neq s$ ) need to be taken into account. You *et al.* (2000, 2003) proposed a cross-sectional and time series model for the LFS unemployment rate estimates. You *et al.* (2000, 2003) only used previous six months of data to predict the current month rate since the LFS sample rotation is based on a six month cycle. Each month, one sixth of the LFS sample is replaced. Thus after six months, the correlation between estimates is weak (see Section 2.1 for the lag correlation coefficients). Let  $y_i = (y_{i1}, \dots, y_{iT})'$ ,  $\theta_i = (\theta_{i1}, \dots, \theta_{iT})'$ , and  $e_i = (e_{i1}, \dots, e_{iT})'$ , where  $T=6$  here. By assuming that  $e_i$  follows a multivariate normal distribution with mean vector 0 and sampling covariance matrix  $\Sigma_i$ , we have

$$y_i \sim N(\theta_i, \Sigma_i), \quad i=1, \dots, m.$$

Thus  $y_i$  is assumed to be design-unbiased for  $\theta_i$ . The sampling covariance matrix  $\Sigma_i$  is unknown in the model. Direct estimates of the sampling covariance matrices are available. It is customary to assume a known sampling variance in area level model-based small area estimation (Rao 2003). For example, the traditional Fay-Herriot model assumes the sampling variance known in the model. Usually



a smoothed estimator of the sampling variance is used. However, recent development on modeling the sampling variance provides an alternative approach to handle the problem of sampling variance; for example, see Wang and Fuller (2003), You and Dick (2004) and You and Chapman (2006). For the unemployment rate estimation, details on smoothing and modeling the sampling variances are given in Section 3.

To borrow strength across regions and time periods, and following You *et al.* (2000, 2003) we can model the true unemployment rate  $\theta_{it}$  by a linear regression model with random effects through auxiliary variables  $x_{it}$ , that is,

$$\theta_{it} = x'_{it}\beta + v_i + u_{it}, \quad i = 1, \dots, m, \quad t = 1, \dots, T, \quad (3)$$

where  $v_i$  is a area random effect assumed to be  $N(0, \sigma_v^2)$  and  $u_{it}$  is a random time and area component. We can further assume that  $u_{it}$  follows a random walk process over time period  $t = 1, \dots, T$ , that is,

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \quad (4)$$

where  $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ . Then  $\text{cov}(u_{it}, u_{is}) = \min(t, s) \sigma_\varepsilon^2$ . The regression vector  $\beta$  and the variance components  $\sigma_v^2$  and  $\sigma_\varepsilon^2$  are unknown in the model and need to be estimated. Combining the model (1), (3) and (4), we obtain a linear mixed model with time components as

$$y_{it} = x'_{it}\beta + v_i + u_{it} + e_{it}, \quad i = 1, \dots, m, \quad t = 1, \dots, T. \quad (5)$$

You *et al.* (2003) showed that the cross-sectional and time series model (5) is better than the Fay-Herriot model in terms of smoothing the direct estimates and CV reduction over the direct estimates for the LFS unemployment rate estimation.

We have used a random walk model for  $u_{it}$ . Rao and Yu (1994) used a stationary autoregressive model for  $u_{it}$ . You *et al.* (2003) showed that the random walk model on  $u_{it}$  had provided better model fit to the unemployment rate estimation than the autoregressive AR(1) model. Datta *et al.* (1999) also used a random walk model to estimate the US unemployment rates at the state level.

### 2.3 Log-linear linking model

However, a limitation of the model (3) is that the linking model for the parameter of interest, the true unemployment rate  $\theta_{it}$ , is a linear model with normal random effects. Since  $\theta_{it}$  is a positive number between 0 and 1, and it is close to 0, the linear linking model with normal random effects may lead to negative estimates for  $\theta_{it}$  for some small areas. To avoid this problem, You, Chen and Gambino (2002) proposed a log-linear linking model for  $\theta_{it}$  as follows:

$$\log(\theta_{it}) = x'_{it}\beta + v_i + u_{it}, \quad i = 1, \dots, m, \quad t = 1, \dots, T. \quad (6)$$

You and Rao (2002) also studied the log-linear linking model for the Fay-Herriot model as the unmatched sampling and linking models with application in the Canadian census undercoverage estimation. The results of You and Rao (2002) and You *et al.* (2002) have shown that the log-linear linking model performs very well in the small area estimation problems. In this paper, we therefore will use the log-linear linking model (6) for the true unemployment rate  $\theta_{it}$ .

## 3. Sampling variance

In general, we can obtain direct sampling variance estimates from survey data. However, these direct estimates are unstable if sample sizes are small. In area level models of small area estimation, the sampling variances are usually assumed to be known (*e.g.*, Fay and Herriot 1979; Datta *et al.* 1999; You and Rao 2002). If the sampling variances are assumed to be known in the model, then reliable (smoothed) estimates of sampling variances are constructed using other auxiliary data and models usually through generalized variance functions (*e.g.*, Dick 1995; Datta *et al.* 1999). In this paper alternatively, we model sampling variance covariance matrix using the direct estimates in a specific way such that we do not need to assume the sampling variances and covariances are known in the model. Thus we simplify the problem of smoothing unknown sampling variance and integrate the sampling variance modeling part into the whole model.

### 3.1 Smoothing sampling covariance matrix

You *et al.* (2000, 2003) used two steps to smooth the sampling covariance matrix. The first step is to obtain a smoothed or common CV for each CMA/UC by computing the average CVs for each CMA/UC over a certain time period, denoted as  $\overline{CV}_i$ , where  $i = 1, 2, \dots, m$ . The second step is to obtain the average lag correlation coefficients over time and all CMA/UCs, denoted as  $\overline{\rho}_{|t-s|}$  for the time lag  $|t-s|$ . This step involves intensive computation. We have used three years (1999 to 2001) of LFS data to compute the smoothed correlation coefficients. We treat the smoothed values over both time and space as the true values in the model. The one-month lag (lag-1) correlation coefficient is obtained as  $\overline{\rho}_1 = 0.48$ , lag-2 correlation coefficient is  $\overline{\rho}_2 = 0.31$ , lag-3 is  $\overline{\rho}_3 = 0.21$ , lag-4 is  $\overline{\rho}_4 = 0.16$ , lag-5 is  $\overline{\rho}_5 = 0.11$  and  $\overline{\rho}_6 = 0.1$ . After lag 6, the lag correlation coefficient is less than 0.1. The lag correlation coefficients decrease as the lag increases. This is consistent with the rotation pattern of the LFS design. Figure 1 shows the smoothed lag correlation coefficients for the LFS unemployment rate estimates.

By using these smoothed CVs and lag correlation coefficients, a smoothed covariance matrix  $\hat{\Sigma}_i$  can be obtained with diagonal elements  $\hat{\sigma}_{ii}^2 = (\overline{CV}_i)^2 y_{it}^2$  and off-diagonal elements  $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|} \hat{\sigma}_{it} \hat{\sigma}_{is}$ . The smoothed  $\hat{\Sigma}_i$  is then treated as known in the model. The study of You *et al.* (2000, 2003) suggests that using the smoothed  $\hat{\Sigma}_i$  in the model can significantly improve the estimates in terms of CV reduction compared to the HB estimates obtained using the direct survey estimates of  $\Sigma_i$  in the model. For more details of the result, see You *et al.* (2003).

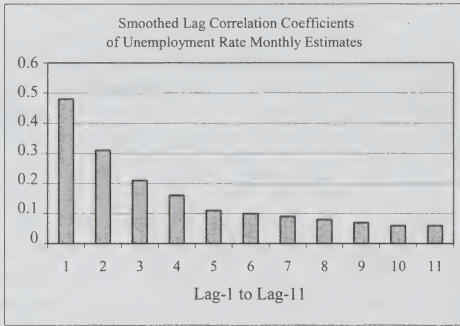


Figure 1 Smoothed unemployment rate lag correlation coefficients

### 3.2 Equal CV modeling approach

The main problem of the method of You *et al.* (2000, 2003) is that the smoothed sampling covariance matrices depend on the direct survey estimates  $y_{it}$ , whereas the  $y_{it}$ 's are not reliable for some small regions. Note that the true sampling variance can be written as  $\sigma_{it}^2 = \theta_{it}^2 (CV_{it})^2$ . Based on the assumption of common CV over time for a given area, You *et al.* (2003) suggested in their concluding remarks to use estimates of the form  $\hat{\sigma}_{it}^2 = \theta_{it}^2 (\overline{CV}_i)^2$  and  $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|} (\hat{\sigma}_{it} \hat{\sigma}_{is})$  for the smoothed variances and covariances respectively. Then the new smoothed sampling covariance matrix  $\hat{\Sigma}_i$  has diagonal elements  $\hat{\sigma}_{ii}^2$  and off-diagonal elements  $\hat{\sigma}_{its}$ . However, under this method, the sampling covariance matrix  $\hat{\Sigma}_i$  becomes unknown in the model, since  $\hat{\sigma}_{it}^2$  and  $\hat{\sigma}_{its}$  depend on the unknown parameters  $\theta_{it}$ , whereas  $\theta_{it}$  is related to a linking model. The advantage of this method is that the model structure of the sampling covariance matrix is clearly specified. This method is better than the smoothing method in the sense that the sampling covariance is clearly specified and not treated as known.

### 3.3 Equal design effects modeling approach

An alternative modeling approach is based on the assumption of common design effects as suggested in

Singh, You and Mantel (2005) and Singh, Folsom and Vaish (2005) to smooth the sampling variance  $\sigma_{it}^2$ . The design effect (deff) for the  $i$ th area at time  $t$  may be approximately written as

$$\text{deff}_{it} = \frac{\sigma_{it}^2}{\theta_{it} (1 - \theta_{it}) / n_{it}},$$

where  $n_{it}$  is the corresponding sample size. Then the sampling variance  $\sigma_{it}^2$  can be written as  $\sigma_{it}^2 = \theta_{it} (1 - \theta_{it}) \cdot \text{deff}_{it} / n_{it}$ . Let  $\tau_{it} = \text{deff}_{it} / n_{it} = \sigma_{it}^2 / (\theta_{it} (1 - \theta_{it}))$ . Then we can estimate  $\tau_{it}$  using the direct estimates of  $\theta_{it}$  and  $\sigma_{it}^2$  as  $\hat{\tau}_{it} = \hat{\sigma}_{it}^2 / (y_{it} (1 - y_{it}))$ . For each area, based on the assumption of a common deff and a common sample size over time, we can obtain a smoothed average factor  $\bar{\tau}_i$  as  $\bar{\tau}_i = \sum_{t=1}^T \hat{\tau}_{it} / T$ . Then a smoothed sampling variance can be obtained as  $\hat{\sigma}_{it}^2 = \theta_{it} (1 - \theta_{it}) \cdot \bar{\tau}_i$ , which again depends on  $\theta_{it}$  as well. The sampling covariance is still in the form of  $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|} (\hat{\sigma}_{it} \hat{\sigma}_{is})$ , as in You *et al.* (2003). Note that  $\bar{\tau}_i$  is a moving average of  $\hat{\tau}_{it}$  over the time period  $T$  in the model. In practice, however, alternatively one may use a longer time series data to obtain more stable estimate of  $\bar{\tau}_i$  for each area if necessary. In this paper, we will use the common design effects model for unemployment rate estimation based on the smoothed moving average factor  $\bar{\tau}_i$  as we borrow information from the past time period  $T$ .

## 4. Hierarchical Bayes inference

In this section, we propose an integrated cross-sectional and time series log-linear model for the unemployment rate estimation. We apply the hierarchical Bayes approach to the model. Estimates of posterior means and posterior variances are obtained by using the Gibbs sampling method.

### 4.1 Integrated hierarchical Bayes model

We now propose the integrated cross-sectional and time series log-linear model in a hierarchical Bayes framework as follows:

- Conditional on  $\theta_i = (\theta_{i1}, \dots, \theta_{iT})'$ ,  $[y_i | \theta_i] \sim \text{ind } N(\theta_i, \Sigma_i(\theta_i))$ ;
- Conditional on  $\beta$ ,  $u_{it}$  and  $\sigma_v^2$ ,  $[\log(\theta_{it}) | \beta, u_{it}, \sigma_v^2] \sim \text{ind } N(x_{it}'\beta + u_{it}, \sigma_v^2)$ ;
- Conditional on  $u_{i,t-1}$  and  $\sigma_\epsilon^2$ ,  $[u_{it} | u_{i,t-1}, \sigma_\epsilon^2] \sim \text{ind } N(u_{i,t-1}, \sigma_\epsilon^2)$ ;
- $\Sigma_i(\theta_i)$  depends on  $\theta_i$  with diagonal elements  $\hat{\sigma}_{ii}^2 = \theta_{it} (1 - \theta_{it}) \cdot \bar{\tau}_i$  and off-diagonal elements  $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|} (\hat{\sigma}_{it} \hat{\sigma}_{is})$ .
- Marginally  $\beta$ ,  $\sigma_v^2$  and  $\sigma_\epsilon^2$  are mutually independent with priors given as  $\beta \propto 1$ ,  $\sigma_v^2 \sim \text{IG}(a_1, b_1)$ , and  $\sigma_\epsilon^2 \sim \text{IG}(a_2, b_2)$ , where IG denotes an inverse gamma

distribution and  $a_1, b_1, a_2, b_2$  are known positive constants and usually set to be very small to reflect our vague knowledge about  $\sigma_v^2$  and  $\sigma_e^2$ .

#### Remarks:

1. The proposed HB model has used a log-linear linking model for the small area parameter of interest  $\theta_{it}$  as suggested in You *et al.* (2002) and You and Rao (2002).
2. The sampling covariance matrix  $\Sigma_i$  is unknown in the model, and it is specified as a function of unknown small area parameter  $\theta_i$  as suggested in You and Rao (2002) and You *et al.* (2003).
3. We have used the assumption of common design effects for small areas as suggested in Singh, You and Mantel (2005).
4. The proposed HB model overcomes the limitations of the model of You *et al.* (2000, 2003) in terms of log-linear modeling and specification of unknown sampling covariance matrix modeling. In particular, we model the unknown sampling covariance matrix through the small area parameters  $\theta_i$  using smoothed estimates of design effects for each area.

We are interested in estimating the true unemployment rate  $\theta_{it}$ , and in particular, the current unemployment rate  $\theta_{iT}$ . In the HB analysis,  $\theta_{iT}$  is estimated by its posterior mean  $E(\theta_{iT} | y)$  and the uncertainty associated with the estimator is measured by the posterior variance  $V(\theta_{iT} | y)$ . We use the Gibbs sampling method (Gelfand and Smith 1990; Gelman and Rubin 1992) to obtain the posterior mean and the posterior variance of  $\theta_{iT}$ .

#### 4.2 Gibbs sampling inference

The Gibbs sampling method is an iterative Markov chain Monte Carlo sampling method to simulate samples from a joint distribution of random variables by sampling from low dimensional densities to make inference about the joint and marginal distributions (Gelfand and Smith 1990). The most prominent application is for inference within a Bayesian framework. In Bayesian inference one is interested in the posterior distribution of the parameters. Assume that  $y_i | \theta$  has conditional density  $f(y_i | \theta)$  for  $i = 1, \dots, n$  and that the prior information about  $\theta = (\theta_1, \dots, \theta_k)'$  is summarized by a prior density  $\pi(\theta)$ . Let  $\pi(\theta | y)$  denote the posterior density of  $\theta$  given the data  $y = (y_1, \dots, y_n)'$ . It may be difficult to sample from  $\pi(\theta | y)$  directly in practice due to the high dimensional integration with respect to  $\theta$ . However, one can use the Gibbs sampler to construct a Markov chain  $\{\theta^{(g)} = (\theta_1^{(g)}, \dots, \theta_k^{(g)})'\}$  with  $\pi(\theta | y)$  as the limiting distribution. For illustration, let  $\theta = (\theta_1, \theta_2)'$ . Starting with an initial set of values  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$ , we generate  $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$  by sampling  $\theta_1^{(g)}$  from  $\pi(\theta_1 | \theta_2^{(g-1)}, y)$

and  $\theta_2^{(g)}$  from  $\pi(\theta_2 | \theta_1^{(g-1)}, y)$ . Under certain regularity conditions,  $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$  converges in distribution to  $\pi(\theta | y)$  as  $g \rightarrow \infty$ . Marginal inference about  $\pi(\theta_i | y)$  can be based on the marginal samples  $\{\theta_i^{(g+k)}; k = 1, 2, \dots\}$  for large  $g$ .

For the proposed integrated HB model, to obtain the posterior estimation of unemployment rate, we implement the Gibbs sampling method by generating samples from the full conditional distributions of the parameters  $\beta$ ,  $\sigma_v^2$  and  $\sigma_e^2$ ,  $u_{it}$  and  $\theta_i$ . These full conditional distributions are given in the Appendix. The distributions of  $\beta$ ,  $\sigma_v^2$  and  $\sigma_e^2$ ,  $u_{it}$  are standard normal or inverse gamma distributions that can be easily sampled. However, the conditional distribution of  $\theta_i$  does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chib and Greenberg 1995) to update  $\theta_i$ . Following You *et al.* (2002) and You and Rao (2002), the full conditional distribution of  $\theta_i$  in the Gibbs sampler can be written as

$$\theta_i | Y, \beta, \sigma_v^2, \sigma_e^2, u \propto h(\theta_i) f(\theta_i),$$

where

$$h(\theta_i) = \left| \sum_i (\theta_i) \right|^{-1} \exp \left\{ -\frac{1}{2} (y_i - \theta_i)' \sum_i^{-1} (y_i - \theta_i) \right\}$$

and

$$f(\theta_i) = \exp \left\{ -\frac{1}{2\sigma_v^2} (\log(\theta_i) - x'_i \beta - u_i)' (\log(\theta_i) - x'_i \beta - u_i) \right\} \cdot \left( \prod_{t=1}^T \frac{1}{\theta_{it}} \right).$$

To update  $\theta_i$ , we proceed as follows:

1. For  $t = 1, \dots, T$ , draw  $\theta_{it}^{(k+1)} \sim \log N(x'_i \beta^{(k+1)} + u_{it}^{(k+1)}, \sigma_v^{2(k+1)})$ , then we have  $\theta_i^{(k+1)} = (\theta_{i1}^{(k+1)}, \dots, \theta_{iT}^{(k+1)})'$ .
2. Compute the rejection probability

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min \left\{ \frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1 \right\}.$$

3. Generate  $\lambda \sim \text{Uniform}(0, 1)$ , if  $\lambda < \alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$ , then accept  $\theta_i^{(k+1)}$ ; otherwise reject  $\theta_i^{(k+1)}$  and set  $\theta_i^{(k+1)} = \theta_i^{(k)}$ .

To implement Gibbs sampling, we follow the recommendation of Gelman and Rubin (1992) and independently run  $L (L > 2)$  parallel chains, each of length  $2d$ . The first  $d$  iterations of each chain are deleted. The convergence monitoring is based on the potential scale reduction factor as suggested in Gelman and Rubin (1992) and adopted by You *et al.* (2003) for estimating  $\theta_{iT}$ . Details are given in You *et al.* (2003). Estimates of the posterior mean  $E(\theta_{iT} | y)$  and the posterior variance  $V(\theta_{iT} | y)$  are obtained based on the samples generated from the Gibbs sampler.



## 5. Application to LFS data

### 5.1 Estimation

We use the 2005 January to June LFS unemployment rate estimates,  $y_{it}$ , in our data analysis. In addition to the direct estimates  $y_{it}$  and the sampling covariance matrices used in the small area models, auxiliary administrative variables are needed in the models. For the unemployment rate estimation, local area employment insurance (EI) monthly beneficiary rate is used as auxiliary data  $x_{it}$  in the model. The beneficiary rate is calculated as the ratio of the number of persons applying EI benefit over the number of persons in the labour force. There are 72 CMA/UCs across Canada. One UC (Miramichi) does not have the EI data. So we consider  $m=71$  CMA/UCs in the model. Within each area, we consider six consecutive monthly estimates  $y_{it}$  from January 2005 to June 2005, so that  $T=6$ . For the January to June 2005 data, the overall average (over 71 CMA/UCs and 6 months) unemployment rate is 0.076, and the overall average EI beneficiary rate is 0.059. For the proposed small area model, the parameter of interest  $\theta_{iT}$  is the true unemployment rate for area  $i$  in June 2005, where  $i = 1, \dots, 71$ . To implement the Gibbs sampler, we have used 10 parallel runs, each of length 2000. The first 1,000 iterations are deleted as “burn-in” periods. The hyperparameters for variance components in the model are set to be 0.0001 to reflect the vague knowledge about  $\sigma_v^2$  and  $\sigma_e^2$ .

We now present the posterior estimates of the unemployment rates under the proposed integrated HB model given in section 4.1 using the Gibbs sampling method. Figure 2 displays the LFS direct estimates and the HB model-based estimates of the June 2005 unemployment rates for the 71 CMA/UCs across Canada. The 71 CMA/UCs appear in the order of population size with the smallest UC (Dawson Creek, BC) on the left and the largest CMA (Toronto, ON) on the right. For the point estimates, the HB estimates leads to moderate smoothing of the direct LFS estimates. For the CMAs with large population sizes and therefore large sample sizes, the direct estimates and the HB estimates are very close to each other as expected, particularly for Toronto, Montreal and Vancouver; for smaller UCs, the direct and HB estimates differ substantially for some regions.

Figure 3 displays the CVs of the estimates. The CV of the HB estimate is taken as the ratio of the square root of the posterior variance and the posterior mean. It is clear from Figure 3 that the direct estimates have very large CVs, particularly for the UCs, the CVs are very large and unstable. The HB estimates have very small and stable CVs compared to the direct estimates. The efficiency gain of the HB estimates is obvious, particularly for the UCs with smaller population sizes. More precisely, we computed the

percent CV reduction for the HB estimators based on the data of June 2005. The percent CV reduction is computed as the difference of the direct CV and HB CV relative to the direct CV. The average CV reduction for UCs is 63% and the CV reduction for CMAs is 35%. As expected, the proposed model has achieved a large CV reduction over the direct estimates, particularly for smaller UCs with smaller sample sizes.

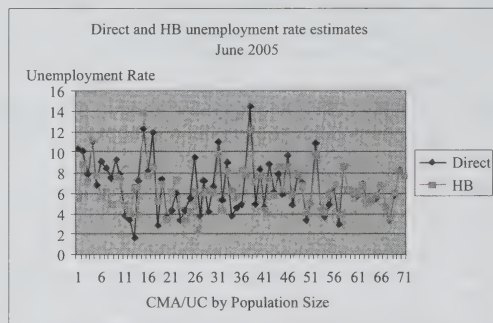


Figure 2 Comparison of direct and HB estimates

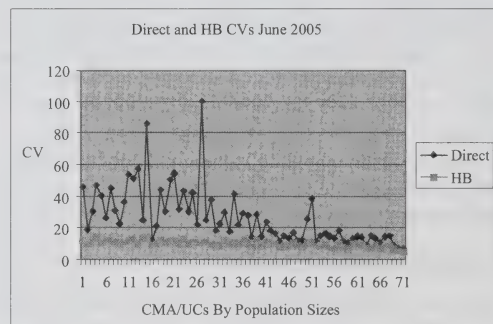


Figure 3 Comparison of direct and HB CVs

### 5.2 Model fit using posterior predictive distribution

To check the overall fit of the proposed model, we use the method of posterior predictive distribution. Let  $y_{rep}$  denote the replicated observation under the model. The posterior predictive distribution of  $y_{rep}$  given the observed data  $y_{obs}$  is defined as

$$f(y_{rep} | y_{obs}) = \int f(y_{rep} | \theta) f(\theta | y_{obs}) d\theta.$$

In this approach, a discrepancy measure  $D(y, \theta)$  that depends on the data  $y$  and the parameter  $\theta$  can be defined and the observed value  $D(y_{obs}, \theta | y_{obs})$  compared to the posterior predictive distribution of  $D(y_{rep}, \theta | y_{obs})$  with any significant difference indicates a model failure. Meng (1994) and Gelman, Carlin, Stern and Rubin (1995) proposed the posterior predictive  $p$ -value as

$$p = P(D(y_{\text{rep}}, \theta) \geq D(y_{\text{obs}}, \theta) | y_{\text{obs}}).$$

This is a natural extension of the usual  $p$ -value in a Bayesian context. If a model fits the observed data, then the two values of the discrepancy measure are similar. In other words, if the given model adequately fits the observed data, then  $D(y_{\text{obs}}, \theta | y_{\text{obs}})$  should be near the central part of the histogram of the  $D(y_{\text{rep}}, \theta | y_{\text{obs}})$  values if  $y_{\text{rep}}$  is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive  $p$ -value is expected to be near 0.5 if the model adequately fits the data. Extreme  $p$ -values (near 0 or 1) suggest poor fit. The posterior predictive  $p$ -value can be estimated as follows: Let  $\theta^*$  represent a draw from the posterior distribution  $f(\theta | y_{\text{obs}})$ , and let  $y_{\text{rep}}^*$  represent a draw from  $f(y_{\text{rep}} | \theta^*)$ . Then marginally  $y_{\text{rep}}^*$  is a sample from the posterior predictive distribution  $f(y_{\text{rep}} | y_{\text{obs}})$ . Computing the  $p$ -value is relatively easy using the simulated values of  $\theta^*$  from the Gibbs sampler. For each simulated value  $\theta^*$ , we can simulate  $y_{\text{rep}}^*$  from the model and compute  $D(y_{\text{rep}}^*, \theta^*)$  and  $D(y_{\text{obs}}, \theta^*)$ . Then the  $p$ -value is estimated by the proportion of times  $D(y_{\text{rep}}^*, \theta^*)$  exceeds  $D(y_{\text{obs}}, \theta^*)$ .

For the proposed HB model, the discrepancy measure used for overall fit is given by  $d(y, \theta) = \sum_{i=1}^n (y_i - \theta_i) \sum_{i=1}^n (y_i - \theta_i)$ . This measure has been used by Datta *et al.* (1999) and You *et al.* (2003). We computed the  $p$ -value by combining the simulated  $\theta^*$  and  $y^*$  from all 10 parallel runs. We obtained an estimated average  $p$ -value about 0.38. Thus we have no indication of lack of overall model fit.

The posterior predictive  $p$ -value model checking has been criticized for being conservative due to the double use of the observed data. The double use of the data can induce unnatural behaviour, as demonstrated by Bayarri and Berger (2000). They proposed alternative model checking  $p$ -value measures, named the partial posterior predictive  $p$ -value and the conditional predictive  $p$ -value. However, their methods are more difficult to implement and interpret (Rao 2002; Sinharay and Stern 2003). As noted in Sinharay and Stern (2003), the posterior predictive  $p$ -value is especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

To compare the proposed model with the model of You *et al.* (2003), we computed the divergence measure of Laud and Ibrahim (1995) based on the posterior predictive distribution. The expected divergence measure of Laud and Ibrahim (1995) is given by  $d(y^*, y_{\text{obs}}) = E(k^{-1} \| y^* - y_{\text{obs}} \|^2 | y_{\text{obs}})$ , where  $k$  is the dimension of  $y_{\text{obs}}$  and  $y^*$  is a sample from the posterior predictive distribution  $f(y | y_{\text{obs}})$ . Between two models, we prefer a model that yields a smaller value of this measure. As in Datta, Day and Maiti (1998) and You *et al.* (2003), we

approximated the divergence measure  $d(y^*, y_{\text{obs}})$  by using the simulated samples from the posterior predictive distribution. Using the Gibbs sampling multiple outputs, we obtained a divergence measure in the range of 8 to 9 for the proposed model, and about 12 to 14 for the model of You *et al.* (2003). Thus the divergence measure suggests a better fit of the proposed integrated HB model for the LFS unemployment rate estimation.

### 5.3 Bias diagnostic using regression analysis

To evaluate the possible bias introduced by the model, we use a simple method of ordinary least squares regression analysis for the direct LFS estimates and the HB model-based estimates. The regression method is suggested by Brown, Chamber, Heady and Heasman (2001). If the model-based estimates are close to the true unemployment rates, then the direct LFS estimators should behave like random variables whose expected values correspond to the values of the model-based estimates. We plot the model-based HB estimates as  $X$  and the direct LFS estimates as  $Y$ , and see how close the regression line is to  $Y = X$ . In terms of regression, basically we fit the regression model  $Y = \alpha X$  to the data and estimate the coefficient  $\alpha$ . Less biased model-based estimates should lead to the value of  $\alpha$  close to 1. For the June 2005 data, let  $Y$  be the direct unemployment rate estimates, and  $X$  be the model-based HB estimates. We obtain the estimated  $\alpha$  value as 1.0207 with standard error 0.0281. Figure 4 shows a scatter plot with the fitted regression line.

The regression result shows no significant difference from  $Y = X$ . Therefore, we conclude that the model-based estimates derived from the proposed model are consistent with the direct LFS estimates with no extra possible bias included. The result may also indicate no evidence of any bias due to possible model misspecification.

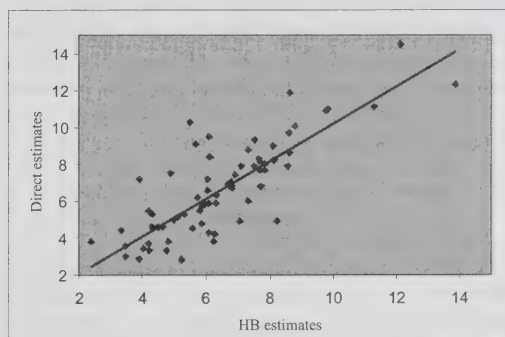


Figure 4 Scatter plot with regression line

## 6. Concluding remarks and future work

In this paper we have reviewed some small area models including the Fay-Herriot model and the cross-sectional and time series model of You *et al.* (2003). In view of the previous work, we have proposed an integrated non-linear cross-sectional and time series model to obtain model-based estimates of unemployment rates for CMA/UCs across Canada using the LFS data. The proposed model overcomes the limitations of the previous work. In particular, we can model the sampling variance as a function of the small area mean by assuming either a common CV for a given area or a common deff for a given area. Our data analysis has shown that the proposed model fits the data quite well. The hierarchical Bayes estimates, based on the model, improve the direct survey estimates significantly in terms of CV reduction, especially for UCs with small population sizes.

We plan to use alternative modeling approach for the sampling variance. Recently You and Dick (2004) and You and Chapman (2006) has used the HB approach to model the sampling variance directly without specifying the form of the sampling variance under the frame of the Fay-Herriot model. The model automatically takes into account the variability of estimating the sampling variances. In particular, You and Dick (2004) applied the model to the census undercoverage estimation problem and obtained efficient HB census undercoverage estimates for small domains across Canada. It will be interesting to adopt the same idea to the cross-sectional and time series model and compare the results with the current work. The purpose of comparison is to establish a reliable and easy-to-implement model for the LFS model-based unemployment rate estimation for small areas.

We plan to produce the model-based estimates for a relative long time period, for example, 24 months from 2004 to 2005. We will compare the 24 months model-based estimates with the 24 months direct estimates, particularly for the large CMAs to study the smoothing effects of the proposed model. The model-based estimates should follow the pattern of direct LFS estimates for large CMAs, which indicates that the smoothing effects on time series effects are reasonable. The purpose is to verify the robustness of the proposed model-based estimates over time.

## Appendix

In the following, we present the full conditional distributions for the Gibbs sampler under the proposed HB model. Let  $Y = (Y'_1, \dots, Y'_m)'$ ,  $X = (X'_1, \dots, X'_m)'$ ,  $\theta = (\theta'_1, \dots, \theta'_m)'$ , and  $u = (u'_1, \dots, u'_m)'$ , with  $Y_i = (y_{i1}, \dots, y_{iT})'$ ,  $X'_i = (x_{i1}, \dots, x_{iT})'$ ,  $\theta'_i = (\theta_{i1}, \dots, \theta_{iT})'$ , and  $u'_i = (u_{i1}, \dots, u_{iT})'$ , we obtain the full conditional distributions as follows:

- $\beta | Y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim N((X'X)^{-1}X'(\log(\theta) - u), \sigma_v^2(X'X)^{-1});$
- $\sigma_v^2 | Y, \beta, \sigma_\varepsilon^2, u, \theta \sim IG\left(\left(a_1 + mT/2, b_1 + \sum_{i=1}^m \sum_{t=1}^T (\log(\theta_{it}) - x'_{it}\beta - u_{it})^2\right)/2\right);$
- $\sigma_\varepsilon^2 | Y, \beta, \sigma_v^2, u, \theta \sim IG\left(\left(a_2 + m(T-1)/2, b_2 + \sum_{i=1}^m \sum_{t=2}^T (u_{it} - u_{i,t-1})^2\right)/2\right);$
- For  $i = 1, \dots, m$ ,  
 $u_{i1} | Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta \sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1} \left(\frac{\log(\theta_{i1}) - x'_{i1}\beta}{\sigma_v^2} + \frac{u_{i2}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$
- For  $i = 1, \dots, m$ , and  $2 \leq t \leq T-1$ ,  
 $u_{it} | Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1}, u_{i,t+1}, \theta \sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1} \left(\frac{\log(\theta_{it}) - x'_{it}\beta}{\sigma_v^2} + \frac{u_{i,t-1} + u_{i,t+1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\right);$
- For  $i = 1, \dots, m$ ,  
 $u_{iT} | Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,T-1}, \theta \sim N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1} \left(\frac{\log(\theta_{iT}) - x'_{iT}\beta}{\sigma_v^2} + \frac{u_{i,T-1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$
- For  $i = 1, \dots, m$ ,  
 $\theta_i | Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - \theta_i)' \Sigma_i^{-1}(y_i - \theta_i)\right\} \times \exp\left\{-\frac{1}{2\sigma_v^2} \sum_{t=1}^T (\log(\theta_{it}) - x'_{it}\beta - u_{it})^2\right\} \left(\prod_{t=1}^T \frac{1}{\theta_{it}}\right).$

## Acknowledgements

The author would like to thank the Editor, the Associate Editor and one referee for their comments and suggestions. This work was partially supported by Statistics Canada Methodology Branch Research Block Fund.



## References

- Bayarri, M.J., and Berger, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, CD-ROM.
- Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.
- Datta, G.S., Day, B. and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā*, 60, 344-362.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.
- Dick, P. (1995). Modeling net undercoverage in the 1991 Canadian Census. *Survey Methodology*, 21, 45-54.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindey, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue No. 71-526.
- Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Laud, P., and Ibrahim, J. (1995). Predictive model selection. *Journal of Royal Statistical Society, Series B*, 57, 247-262.
- Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.
- Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical methods Conference proceedings, Washington, D.C., www.fcsm.gov*.
- Singh, A., You, Y. and Mantel, H. (2005). Use of generalized design effects for variance function modeling in small area estimation from survey data. Presentation at the 2005 Statistical Society of Canada Annual Meeting, Regina, SK.
- Sinharay, S., and Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.
- Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.
- You, Y., Chen, E. and Gambino, G. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics*, Alexandria, VA: American Statistical Association. 3883-3888.
- You, Y., and Dick, P. (2004). Hierarchical Bayes small area inference to the 2001 census undercoverage estimation. *2004 Proceedings of the American Statistical Association, Section on Government Statistics*, Alexandria, VA: American Statistical Association, 1836-1840.
- You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.
- You, Y., Rao, J.N.K. and Gambino, J. (2000). Hierarchical Bayes estimation of unemployment rates for sub-provincial regions using cross-sectional and time series data. *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*, 160-165.
- You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.



# Small area estimation under a restriction

Junyuan Wang, Wayne A. Fuller and Yongming Qu<sup>1</sup>

## Abstract

Small area prediction based on random effects, called EBLUP, is a procedure for constructing estimates for small geographical areas or small subpopulations using existing survey data. The total of the small area predictors is often forced to equal the direct survey estimate and such predictors are said to be calibrated. Several calibrated predictors are reviewed and a criterion that unifies the derivation of these calibrated predictors is presented. The predictor that is the unique best linear unbiased predictor under the criterion is derived and the mean square error of the calibrated predictors is discussed. Implicit in the imposition of the restriction is the possibility that the small area model is misspecified and the predictors are biased. Augmented models with one additional explanatory variable for which the usual small area predictors achieve the self-calibrated property are considered. Simulations demonstrate that calibrated predictors have slightly smaller bias compared to those of the usual EBLUP predictor. However, if the bias is a concern, a better approach is to use an augmented model with an added auxiliary variable that is a function of area size. In the simulation, the predictors based on the augmented model had smaller MSE than EBLUP when the incorrect model was used for prediction. Furthermore, there was a very small increase in MSE relative to EBLUP if the auxiliary variable was added to the correct model.

**Key Words:** Components-of-variance model; Best linear unbiased prediction; Calibration; Design consistent; Mixed linear models.

## 1. Introduction

There are situations in which it is desirable to derive reliable estimators for small geographical areas or small subpopulations from existing survey data. However, sample sizes for the areas may be such that the usual survey estimators yield unacceptably large standard errors. This makes it reasonable to use a model-based estimator. See Rao (2003) for a complete discussion of small area estimation.

A model for small area estimation is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + b_i, \quad (1)$$

$$Y_i = y_i + e_i, \quad i = 1, \dots, n, \quad (2)$$

where  $y_i$  are unobservable small area means,  $Y_i$  are observable survey estimators,  $\mathbf{x}_i'$  are known vectors,  $\boldsymbol{\beta}$  is the vector of regression parameters,  $b_i$  are independent and identically distributed random variables with  $E(b_i) = 0$  and  $V(b_i) = \sigma_b^2$ , and  $e_i$  are sampling errors with  $E(e_i | y_i) = 0$  and  $V(e_i | y_i) = \sigma_{ei}^2$ . Combining (1) and (2), we obtain

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + b_i + e_i, \quad i = 1, \dots, n, \quad (3)$$

which is a special case of the mixed linear model.

Assuming the variance components  $\sigma_b^2$  and  $\sigma_{ei}^2$  to be known, the best linear unbiased estimator of  $\boldsymbol{\beta}$  is

$$\hat{\boldsymbol{\beta}} = [\mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{X}]^{-1} \mathbf{X}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} \\ = \left[ \sum_{i=1}^n (\sigma_b^2 + \sigma_{ei}^2)^{-1} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[ \sum_{i=1}^n (\sigma_b^2 + \sigma_{ei}^2)^{-1} \mathbf{x}_i Y_i \right], \quad (4)$$

where  $\mathbf{X}' = (\mathbf{x}_1', \dots, \mathbf{x}_n')$ ,  $\mathbf{Y}' = (Y_1, \dots, Y_n)$ , and  $\boldsymbol{\Sigma} = \text{Var}(\mathbf{Y}) = \text{diag}(\sigma_b^2 + \sigma_{e1}^2, \dots, \sigma_b^2 + \sigma_{en}^2)$ . Furthermore, the best linear unbiased predictor (BLUP) of  $y_i$  is

$$\hat{y}_i^H = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + \gamma_i (Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}), \quad (5)$$

where

$$\gamma_i = (\sigma_b^2 + \sigma_{ei}^2)^{-1} \sigma_b^2. \quad (6)$$

See Henderson (1963) and Rao (2003). When the variance components are unknown, we replace the variance components in (4) and (6) with estimators to obtain  $\hat{y}_i^H$ , the empirical BLUP or EBLUP.

The survey estimator of the total of all survey areas is often judged to be of adequate precision. In such cases, the practitioner may prefer to use the design consistent estimator of the total and to require that the weighted sum of the small area predictors equal the design consistent estimator. Thus, it is desirable to have small area predictors  $\hat{y}_i$  that satisfy

$$\sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i, \quad (7)$$

where  $\omega_i$  are sampling weights such that  $\sum_{i=1}^n \omega_i Y_i$  is a design consistent estimator of the total (or mean). A number of procedures have been suggested for constructing predictors to satisfy (7). Such procedures are often called "benchmarking" or "calibration", e.g., Mantel, Singh and Barreau (1993) and You and Rao (2003).

To review such procedures, let  $\hat{\mathbf{y}}^H = (\hat{y}_1^H, \dots, \hat{y}_n^H)'$  denote the BLUP predictor of  $\mathbf{y} = (y_1, \dots, y_n)'$  defined in (5), where

1. Junyuan Wang, Director of Biostatistics, Global Biostatistics and Programming, Wyeth Research, 35 Cambridge Park Drive, Cambridge, MA 02140; Wayne A. Fuller, Professor, Department of Statistics, Iowa State University, Ames, IA 50011; Yongming Qu, Sr. Research Scientist, Eli Lilly and Company, Lilly Corporation Center, Indianapolis, IN 26285.



$$\tilde{\mathbf{y}}^H = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{b}}, \quad (8)$$

$\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{b}}$  are any solutions to

$$\begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{X} & \mathbf{X}'\boldsymbol{\Sigma}_e^{-1} \\ \boldsymbol{\Sigma}_e^{-1}\mathbf{X} & \boldsymbol{\Sigma}_e^{-1} + \boldsymbol{\Sigma}_b^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \\ \boldsymbol{\Sigma}_e^{-1}\mathbf{Y} \end{bmatrix}, \quad (9)$$

$\boldsymbol{\Sigma}_b = \sigma_b^2 \mathbf{I}_n$ , and  $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_{e1}^2, \dots, \sigma_{en}^2)$ . Equation (9) is called the mixed model equation. Finding a solution to a mixed model equation (9) is equivalent to finding a solution to the minimization problem

$$\min_{\boldsymbol{\beta}, \mathbf{b}} \{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b})' \boldsymbol{\Sigma}_e^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{b}) + \mathbf{b}' \boldsymbol{\Sigma}_b^{-1} \mathbf{b}\}. \quad (10)$$

Pfeffermann and Barnard (1991) proposed the modified predictor

$$\hat{\mathbf{y}}^{\text{PB}} = \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{PB}} + \hat{\mathbf{b}}^{\text{PB}}, \quad (11)$$

where  $\hat{\boldsymbol{\beta}}^{\text{PB}}$  and  $\hat{\mathbf{b}}^{\text{PB}}$  are any solutions to the minimization problem (10) with  $\boldsymbol{\beta}$  and  $\mathbf{b}$  subject to the constraint

$$\sum_{i=1}^n \omega_i (\mathbf{x}_i' \boldsymbol{\beta} + b_i) = \sum_{i=1}^n \omega_i Y_i. \quad (12)$$

This leads to the predictor

$$\hat{\mathbf{y}}_i^{\text{PB}} = \tilde{\mathbf{y}}_i^H + [\text{Var}(\tilde{\mathbf{y}})]^{-1} \text{cov}(\tilde{\mathbf{y}}_i^H, \tilde{\mathbf{y}}) \left[ \sum_{j=1}^n \omega_j Y_j - \tilde{\mathbf{y}} \right], \quad (13)$$

where  $\tilde{\mathbf{y}} = \sum_{i=1}^n \omega_i \tilde{\mathbf{y}}_i^H$ ,  $\text{cov}(\tilde{\mathbf{y}}_i^H, \tilde{\mathbf{y}}) = \omega_i \gamma_i \sigma_{ei}^2 + \sum_{j=1}^n \omega_j (1 - \gamma_i)(1 - \gamma_j) \mathbf{x}_i' \mathbf{V}(\hat{\boldsymbol{\beta}}) \mathbf{x}_j$ , and  $\text{Var}(\tilde{\mathbf{y}}) = \sum_{i=1}^n \omega_i \text{cov}(\tilde{\mathbf{y}}_i^H, \tilde{\mathbf{y}})$ .

Isaki, Tsay and Fuller (2000) imposed the restriction by a procedure that, approximately, constructs the best predictors of  $n-1$  quantities that are uncorrelated with  $\sum_{i=1}^n \omega_i Y_i$ . After some matrix operations, the Isaki-Tsay-Fuller (ITF) predictor can be rewritten as

$$\hat{\mathbf{y}}_i^{\text{ITF}} = \hat{\mathbf{y}}_i^H + \left[ \sum_{j=1}^n \omega_j^2 \widehat{\text{Var}}(Y_j) \right]^{-1} \omega_i \widehat{\text{Var}}(Y_i) \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{\mathbf{y}}_j^H \right), \quad (14)$$

where  $\widehat{\text{Var}}(Y_i)$  is an estimator of  $\sigma_b^2 + \sigma_{ei}^2$ .

Note that the Pfeffermann-Barnard (PB) predictor (13), and the ITF predictor (14) have the form

$$\hat{\mathbf{y}}_i^a = \hat{\mathbf{y}}_i + a_i \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{\mathbf{y}}_j \right), \quad (15)$$

where  $\sum_{i=1}^n \omega_i a_i = 1$ . In other words, we may consider imposing restriction (7) to be an adjustment problem. To make an adjusted predictor  $\hat{\mathbf{y}}_i^a$  satisfy (7), the difference  $\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{\mathbf{y}}_j$  is allocated to small area predictor  $\hat{\mathbf{y}}_i$  using  $a_i$ .

Using the unit level model, You and Rao (2002) proposed an estimator of  $\boldsymbol{\beta}$  such that the resulting predictors satisfy (7). They called such predictors self-calibrated. Applying their procedure to the area model (3), we have

$$\hat{\mathbf{y}}_i^{\text{YR}} = \hat{\gamma}_i Y_i + (1 - \hat{\gamma}_i) \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\text{YR}}, \quad (16)$$

where

$$\hat{\boldsymbol{\beta}}_{\text{YR}} = \left[ \sum_{i=1}^n \omega_i (1 - \hat{\gamma}_i) \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \sum_{i=1}^n \omega_i (1 - \hat{\gamma}_i) \mathbf{x}_i Y_i. \quad (17)$$

Any predictor that has the self-calibrated property, such as the You and Rao (YR) predictor (16), is a predictor of the form (15) since the difference  $\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{\mathbf{y}}_j$  is equal to zero.

We will derive the "best" predictor of the form (15) in Section 2. The results will lead to a unifying view of several BLUP based predictors. In Section 3, we propose an alternative approach that has the self-calibrated property. We will briefly discuss the mean square error (MSE) in Section 4 and use simulation studies to compare the predictors in Section 5. Conclusions and discussion will be given in Section 6.

## 2. "Best" linear unbiased predictor under a restriction

To find the "best" linear unbiased predictor for  $\mathbf{y}$  that satisfies restriction (7), we first assume the parameters for the variance components are known. According to Lemma 1 of Pfeffermann and Barnard (1991), it is impossible to compare predictors that satisfy restriction (7) component-by-component to find the best one. Therefore, some kind of overall criterion is required. A natural criterion is

$$Q(\hat{\mathbf{y}}^a) = \sum_{i=1}^n \varphi_i E(\hat{\mathbf{y}}_i^a - y_i)^2, \quad (18)$$

where the  $\varphi_i$ ,  $i = 1, \dots, n$  are a chosen set of positive weights.

**Theorem 1.** Assume the random effects model

$$Y_i = \mathbf{x}_i' \boldsymbol{\beta} + b_i + e_i, \quad i = 1, \dots, n,$$

where the  $b_i$  have independent identical distributions with mean zero and variance  $\sigma_b^2$ , the  $e_i$  have independent distributions with mean zero and variance  $\sigma_{ei}^2$ , and  $\mathbf{b} = (b_1, \dots, b_n)'$  is independent of  $\mathbf{e} = (e_1, \dots, e_n)'$ . Assume  $\sigma_b^2$  and  $\sigma_{ei}^2$  are known, and  $\boldsymbol{\beta}$  is unknown. Let  $\tilde{\mathbf{y}}_i^H$  be the BLUP of  $y_i$  defined in (5). Let

$$\hat{y}_i^a = \hat{y}_i^H + \tilde{a}_i \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H \right), \quad (19)$$

where  $\tilde{a}_i = (\sum_{i=1}^n \phi_i^{-1} \omega_i^{-1})^{-1} \phi_i^{-1} \omega_i$  and  $\omega_i$  are the fixed weights of (7). Then  $\hat{y}^a = (\hat{y}_1^a, \dots, \hat{y}_n^a)'$  is the unique predictor among all linear unbiased predictors that satisfy (7) and minimize criterion (18).

**Proof:** See Appendix A.

**Remark 1.** When the variance components are unknown, we replace the variance components in (6) with suitable estimators to obtain the empirical BLUP or EBLUP, denoted by  $\hat{y}_i^H$ . Thus, we have the modified predictor

$$\hat{y}_i^a = \hat{y}_i^H + \tilde{a}_i \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H \right). \quad (20)$$

**Remark 2.** The criterion (18) defines a “loss function”, where the choice of the weights  $\phi_i$  depends on the problem under consideration. For example, a statistician can decide to assign higher weights to “more important” areas and lower weights to “less important” areas. Often,  $\phi_i$  is function of the variance components. In some cases, one can choose  $\phi_i$  so that the derived predictors have certain desirable properties. For example,  $\phi_i = [\widehat{\text{Var}}(Y_i)]^{-1}$  gives the ITF predictors, which are BLUP (in the traditional sense) in the subspace that is orthogonal to  $(\omega_1, \dots, \omega_n)'$  in the space spanned by  $Y$ .

**Remark 3.** If  $\phi_i = \omega_i [\text{cov}(\tilde{y}_i^H, \tilde{y})]^{-1}$ , where  $\tilde{y} = \sum_{j=1}^n \omega_j \hat{y}_j^H$ , we have the predictor (13) derived by Pfeffermann and Barnard (1991). When  $\phi_i = [\widehat{\text{Var}}(\hat{y}_j^H)]^{-1}$ , we have the predictor used by Battese, Harter, and Fuller (1988).

### 3. An alternative way to impose the restriction

We have discussed a family of predictors in which the total for the small area predictors is equal to the total of the direct survey estimates. Implicit in the imposition of restriction (7) is the possibility that the small area predictor of the total is biased due to a misspecified model (3). In practical applications, model misspecification is a valid concern since the true mechanism that generates  $Y$  is unknown.

A common misspecification occurs when the explanatory variables used in the model are not the same as the ones that generated  $Y$ . Thus, the direction of the overall bias may not be the same as the direction of the bias for a particular small area. In this case, the predictors of form (15) may increase the bias for some small areas compared to the bias before adjustment. Mantel *et al.* (1993) concluded that “Generally the effect of benchmarking here is a slight improvement in

the overall bias at the cost of some deterioration with respect to the other evaluation measures”.

Since the bias is nonzero if there is nonzero correlation between  $\omega_i$  and  $(Y_i - \hat{y}_i)$ , the bias can be reduced by including  $\omega_i$  in the model. That is, for a given model, one approach is to use the augmented model

$$Y = X_1 \beta + b + e, \quad (21)$$

where  $X_1 = (X, \omega)$  and  $\omega = (\omega_1, \dots, \omega_n)'$ , to obtain the BLUP or EBLUP. With  $\omega$  in the model, the adjustment needed to meet restriction (7) will often be much smaller than the adjustment for the model without  $\omega$ .

Using an augmented model approach, we can go one step further and construct predictors that satisfy restriction (7). First, assume the variances  $\sigma_{ei}^2$  are known. Note that

$$\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_j^H = \sum_{i=1}^n \omega_i (1 - \gamma_i) (Y_i - x_i' \hat{\beta}) \quad (22)$$

and  $\omega_i (1 - \gamma_i) \text{Var}(Y_i) = \omega_i \sigma_{ei}^2$ . Using the theory of the linear model, we can show that the predictor constructed with the augmented model

$$Y = X_2 \beta + b + e, \quad (23)$$

where  $X_2 = (X, \omega_e)$  and  $\omega_e = (\omega_1 \sigma_{e1}^2, \dots, \omega_n \sigma_{en}^2)'$ , has the self-calibrated property when the generalized least squares (GLS) estimator of  $\beta$  is used. Note that this approach gives a predictor that is different than the You-Rao predictor (16).

If the  $\sigma_{ei}^2$  are unknown, we replace  $\sigma_{ei}^2$  in  $\omega_e$  with its estimator  $\hat{\sigma}_{ei}^2$ . As long as the  $\hat{\sigma}_{ei}^2$  in  $\omega_e$  is the same as the  $\hat{\sigma}_{ei}^2$  used in constructing  $\hat{y}_i$ , the predictors have the self-calibrated property. If the  $\sigma_{ei}^2$  has the form  $\sigma_e^2 f(u_i)$ , where  $\sigma_e^2$  is unknown, but  $u_i$  and  $f(\cdot)$  are known, one can construct the variables for model (23) using  $\omega_e = (\omega_1 f(u_1), \dots, \omega_n f(u_n))'$  without estimating  $\sigma_e^2$ . For example, if  $\sigma_{ei}^2 = m_i^{-1} \sigma_e^2$ , the  $\omega_e$  for model (23) is  $(m_1^{-1} \omega_1, \dots, m_n^{-1} \omega_n)'$ .

### 4. The MSE of the modified predictors

One can show that any predictor of the form (15) can be written as

$$\tilde{y}^a = Y - C_a^{-1} B C_a (I_n - \Gamma) (Y - X \hat{\beta}) \quad (24)$$

by letting  $C_a = A_a T$ , where

$$A_a = \begin{pmatrix} 1 & 0'_{n-1} \\ -a_{n-1} & I_{n-1} \end{pmatrix},$$

$$a_{n-1} = (a_2, \dots, a_n)'$$

Therefore, the estimator for the variance of  $\tilde{y}^a$  defined in (14) proposed in Isaki *et al.* (2000) can be used to estimate the MSE of any predictor of the form (15). Often, the MSE of an adjusted predictor is close to the MSE of the predictor before the adjustment.

The augmented model (23) has the self-calibrated property, thus the MSE can be estimated using the formula for usual EBLUP predictors.

## 5. Simulation study

### 5.1 Simulation setup

To study the empirical properties of small area predictors described in Section 2 and 3, we use data designed to simulate a large national survey in which state estimates are of interest. Table 1 contains the approximate populations of the 50 states of the United States in the year 2000. The sample sizes  $m_i$  given in the table are approximately proportional to the square roots of the state populations.

**Table 1 Population and sample sizes for the simulation**

State	Population (in 1,000)	Sample size ( $m_i$ )	State	Population (in 1,000)	Sample size ( $m_i$ )
1	33,640	58	26	4,000	20
2	21,160	46	27	3,610	19
3	19,360	44	28	3,240	18
4	16,000	40	29	3,240	18
5	12,250	35	30	2,890	17
6	12,250	35	31	2,890	17
7	11,560	34	32	2,560	16
8	10,240	32	33	2,560	16
9	8,410	29	34	2,250	15
10	8,410	29	35	1,960	14
11	7,840	28	36	1,690	13
12	7,290	27	37	1,690	13
13	6,250	25	38	1,690	13
14	6,250	25	39	1,210	11
15	5,760	24	40	1,210	11
16	5,760	24	41	1,210	11
17	5,760	24	42	1,210	11
18	5,290	23	43	1,000	10
19	5,290	23	44	810	9
20	5,290	23	45	810	9
21	4,840	22	46	810	9
22	4,410	21	47	640	8
23	4,410	21	48	640	8
24	4,410	21	49	640	8
25	4,000	20	50	490	7

A total of 10,000 samples were generated. Each sample in the simulation study was composed of observations generated from model

$$Y_{ij} = \mathbf{x}'_i \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad (25)$$

where  $\mathbf{x}'_i = (1, z_i)$ ,  $\boldsymbol{\beta}' = (6.0, 3.0)$ ,  $z_i = \text{Pop}_i^{0.2} - \overline{\text{Pop}}^{0.2}$ ,  $\text{Pop}_i$  is the population of state  $i$  in millions,  $\overline{\text{Pop}}^{0.2}$  is the

mean of  $\text{Pop}_i^{0.2}$ ,  $b_i \sim NI(0, 1)$ , and  $\varepsilon_{ij} \sim NI(0, 16)$ . The  $b_i$ 's and  $\varepsilon_{ij}$ 's are independent. The model for the state observations becomes

$$Y_i = \mathbf{x}'_i \boldsymbol{\beta} + b_i + e_i, \quad (26)$$

where  $Y_i = m_i^{-1} \sum_{j=1}^{m_i} Y_{ij}$  and  $e_i = m_i^{-1} \sum_{j=1}^{m_i} \varepsilon_{ij}$ . With the sample sizes given in Table 1, the  $\gamma_i$  defined in (6) is 0.784 for the largest state (California) and 0.304 for the smallest state (Wyoming).

To investigate the performance of five predictors, EBLUP, Pfeffermann-Bernard (PB), Isaki-Tsay-Fuller (ITF), You-Rao (YR), and augmented model (23) (AUG2), two estimation models were used. The first model is a misspecified model with  $\mathbf{x}'_i = 1$  in the notation of (26). This model is called model (A). Correspondingly, the data generating model (26) with  $\mathbf{x}'_i = (1, z_i)$  is called model (B).

Following the method outlined in Wang and Fuller (2003), the estimator of  $\sigma_b^2$  is

$$\hat{\sigma}_b^2 = \max\{0.5[\hat{\nu}(\hat{\sigma}_b^2)]^{0.5}, \hat{\sigma}_b^2\}, \quad (27)$$

where

$$\hat{\sigma}_b^2 = \sum_{i=1}^{50} c_i \left[ \frac{50}{50-k} (Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{OLS}})^2 - \hat{\sigma}_{ei}^2 \right], \quad (28)$$

$$\hat{\nu}(\hat{\sigma}_b^2) = \sum_{i=1}^{50} c_i^2 \left\{ \left[ \frac{50}{50-k} (Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{OLS}})^2 - \hat{\sigma}_{ei}^2 \right] - \hat{\sigma}_b^2 \right\}^2, \quad (29)$$

$c_i = (\sum_{i=1}^{50} m_i^{0.5})^{-1} m_i^{0.5}$ ,  $\hat{\boldsymbol{\beta}}_{\text{OLS}}$  is the ordinary least squares estimator of the regression coefficient of  $Y_i$  on  $\mathbf{x}_i$ ,  $k$  is the dimension of the vector  $\mathbf{x}_i$ ,  $\hat{\sigma}_{ei}^2 = m_i^{-1} s_i^2$ , and  $s_i^2 = (m_i - 1)^{-1} \sum_{j=1}^{m_i} (Y_{ij} - Y_i)^2$  is the sample variance of area  $i$ .

The EBLUP predictor is

$$\hat{y}_i = \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{GLS}} + \hat{\gamma}_i (Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_{\text{GLS}}), \quad (30)$$

where

$$\hat{\gamma}_i = (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \hat{\sigma}_b^2, \quad (31)$$

and

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} = \left[ \sum_{i=1}^n (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \mathbf{x}_i \mathbf{x}_i' \right]^{-1} \left[ \sum_{i=1}^n (\hat{\sigma}_b^2 + \hat{\sigma}_{ei}^2)^{-1} \mathbf{x}_i Y_i \right] \quad (32)$$

is the generalized least squares estimator of  $\boldsymbol{\beta}$ . The restriction considered is  $\sum_{i=1}^{50} \omega_i \hat{y}_i^a = \sum_{i=1}^{50} \omega_i Y_i$ , where

$$\omega_i = \left( \sum_{i=1}^{50} \text{Pop}_i \right)^{-1} \text{Pop}_i, \quad (33)$$

With EBLUP predictor, PB and ITF are derived using (13) and (14). The YR predictor is derived use (16) with  $\hat{\gamma}_i$  defined in (31) and  $\hat{\boldsymbol{\beta}}_{\text{YR}}$  defined in (17). The AUG2



predictor is of the form (30) with  $\mathbf{x}'_i = (1, \omega_i \hat{\sigma}_{ei}^2)$  under the augmented model (A) and  $\mathbf{x}'_i = (1, z_i, \omega_i \hat{\sigma}_{ei}^2)$  under the augmented model (B).

For each of the 10 predictors, the criterion

$$Q(\hat{y}) = 0.02 \sum_{i=1}^{50} \varphi_i (\hat{y}_i - y_i)^2 \tag{34}$$

was calculated, where  $\varphi_i = (\gamma_i m_i^{-1} \sigma_e^2)^{-1}$ . Note that  $\varphi_i^{-1}$  is the variance of the predictor of  $y_i$  constructed with known parameters.

5.2 Simulation results

The estimator of  $\sigma_b^2$  constructed under model (B) has a Monte Carlo mean of 1.001 with a standard deviation of 0.386. When model (A) is used for estimation, the Monte Carlo mean of the estimator of  $\sigma_b^2$  is 1.720 with a standard deviation of 0.521. The mean square of  $3z_i$  is 0.636. Thus, the estimation procedure using model (A) incorporates much of the fixed area effect of model (B) into the random effect. With a bigger  $\hat{\sigma}_b^2$ , the  $\hat{\gamma}_i$  is bigger and a higher proportion of  $Y_i$  was used to construct the predictors, which partially offsets the impact of model misspecification. Under augmented model (A), the estimator of  $\sigma_b^2$  is 1.197, i.e., a much smaller portion of the fixed area effect of model (B) is incorporated into the random effect.

Table 2 contains some summary statistics for predictors based on 10,000 simulated samples. The empirical bias in  $\sum \omega_i (Y_i - \hat{y}_i)$  is zero for all predictors when model (B) or its augmented model is used because the prediction model matches the data-generating model. The difference  $\sum \omega_i (Y_i - \hat{y}_i)$  has a simulated standard deviation of 0.022 for the usual EBLUP predictor. The standard deviations of  $\sum \omega_i (Y_i - \hat{y}_i)$  are 0 for the other four predictors because the predictors satisfy the restriction  $\sum_{i=1}^{50} \omega_i \hat{y}_i = \sum_{i=1}^{50} \omega_i Y_i$ .

Table 2 Monte Carlo properties of small area predictors (average of 10,000 samples generated by model B)

Quantity	EBLUP	PB	ITF	YR	AUG2
Predictor constructed under model (A)					
$\sum \omega_i (Y_i - \hat{y}_i)$ Mean	-0.100	0.000	0.000	0.000	0.000
(SD)	(0.027)	(0.000)	(0.000)	(0.000)	(0.000)
$Q(\hat{y})$ Mean	1.438	1.446	1.419	1.558	1.298
Predictor constructed under model (B)					
$\sum \omega_i (Y_i - \hat{y}_i)$ Mean	-0.000	0.000	0.000	0.000	0.000
(SD)	(0.022)	(0.000)	(0.000)	(0.000)	(0.000)
$Q(\hat{y})$ Mean	1.203	1.202	1.202	1.208	1.219

Prediction based on model (A), or its augmented model, is biased because the data generation model (B) contains a function of the population size. The simulated mean of the weighted difference  $\sum \omega_i (Y_i - \hat{y}_{iA})$  is -0.100, where  $\hat{y}_{iA}$  is the EBLUP predictor. The *t*-statistic for the weighted bias is

-3.70. The simulated variance of the weighted mean of the predictions is

$$V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} = 0.060.$$

The estimated mean square error of the model (A) prediction of  $\sum \omega_i y_i$  for data generated by model (B) is

$$\begin{aligned} \text{MSE} \left\{ \sum_{i=1}^{50} \omega_i (\hat{y}_{iA} - y_i) \right\} &= V \left\{ \sum_{i=1}^{50} \omega_i \hat{y}_{iA} \right\} + \text{Bias}^2 \\ &= 0.060 + (-0.100)^2 = 0.070. \end{aligned} \tag{35}$$

The variance of  $\sum \omega_i Y_i$  is

$$V \left\{ \sum_{i=1}^{50} \omega_i Y_i \right\} = \sum_{i=1}^{50} \omega_i^2 (\sigma_b^2 + m_i^{-1} \sigma_e^2) = 0.0622. \tag{36}$$

Thus, the use of  $\sum \omega_i Y_i$  as the estimator of  $\sum \omega_i y_i$  is about 12.5% more efficient than the predictor  $\sum \omega_i \hat{y}_{iA}$  based on the model (A). Due to calibration, the MSE of the four predictors (PB, ITF, YR, and AUG2) of  $\sum \omega_i y_i$  have the same MSE as the MSE of the directly estimated mean  $\sum \omega_i Y_i$ . The squared bias would be a much larger proportion of the mean square error if there were more small areas.

The value of criterion  $Q$  for EBLUP is 1.20 under model (B). Thus, estimation of the parameters increased the average variance of the predictors about 20% relative to the use of known parameters. If the predictions are made using the known  $\sigma_{ei}^2$ , the value of criterion  $Q$  for EBLUP is 1.06. Therefore, estimation of  $\hat{\sigma}_{ei}^2$  contributes the most to the increase in the variability. Because the bias is zero when model (B) is used for estimation, the adjustments that the restricted predictors make are small. Consequently, the adjustment predictors give criteria values very similar to those of the unadjusted predictors. The YR predictors have slightly larger criterion values than the corresponding PB and ITF predictors because the YR predictor uses an inefficient estimator of  $\beta$ . Predictors based on the augmented model have slightly larger criterion value  $Q$  because the model has a redundant variable. The less than 2% increase in  $Q$  is on the order of  $n^{-1}$ , which is the expected loss from adding an unnecessary parameter in a least square prediction.

The value of criterion  $Q$  for the EBLUP under model (A) is 1.438 compared to 1.203 for the EBLUP under model (B). This is the penalty due to the model misspecification. Among adjustment procedures based on model (A), the ITF procedure has the smallest value for  $Q$ . There is little difference among the PB, ITF, and the EBLUP predictors. The  $Q$  for the You-Rao procedure is about 8% larger than that for the EBLUP. The predictor based on the augmented

model has a  $Q$  about 11% smaller than that of the EBLUP predictor. In a sense, the augmented model is less misspecified.

The augmentation approach not only calibrates the small area predictors, it also reduces the bias at the area level when the model is misspecified. Tables 3 and 4 contain Monte Carlo properties of predictors for some selected areas. In the tables, the estimated biases are normalized by  $(\gamma_i m_i^{-1} \sigma_e^2)^{0.5}$ , the square root of the MSE of the BLUP predictor with known parameters, and the estimated MSE's are normalized by  $\gamma_i m_i^{-1} \sigma_e^2$ . When the correct model (B) is used, the Monte Carlo bias at the individual area is close to zero. See Table 3. Also there is little difference in the MSE's of the different procedures, with the augmented model having slightly (less than 2%) larger MSE. Again, Table 3 shows that the calibration process has little effect on the MSE of the area predictors in compared to the EBLUP predictor.

**Table 3**  
Monte Carlo Properties of individual area predictors using versions of model (B) (10,000 samples generated by model B)

State	Quantity	EBLUP	PB	ITF	YR	AUG2
1	Bias	0.011	0.012	0.012	0.013	0.011
	MSE	1.100	1.101	1.105	1.104	1.120
2	Bias	0.000	0.001	0.001	0.002	0.001
	MSE	1.072	1.072	1.072	1.076	1.092
14	Bias	-0.001	-0.001	-0.001	-0.001	-0.001
	MSE	1.058	1.058	1.058	1.058	1.074
26	Bias	0.015	0.015	0.015	0.015	0.018
	MSE	1.078	1.077	1.078	1.079	1.092
38	Bias	-0.005	-0.005	-0.005	-0.006	-0.003
	MSE	1.123	1.122	1.122	1.132	1.135
50	Bias	0.012	0.012	0.012	0.009	0.014
	MSE	1.222	1.222	1.222	1.247	1.246

If a misspecified model, such as model (A) is used, the bias in the sum of the EBLUP predictors as an estimator of the total is negative for the example because the states with a negative bias have large  $\omega_i$ . See Table 4. The adjustment procedures such as PB or ITF allocate the bias to all the small areas. Thus, the adjustment reduces the negative bias of area predictors with large negative bias and increases the positive bias of predictors with large positive bias. This results in a smaller MSE for larger states and a slightly larger MSE for smaller states. The YR predictor has larger bias than the ITF predictor. On the other hand, predictors constructed with  $\mathbf{x}_i' = (1, \omega_i \hat{\sigma}_e^2)$ , i.e., the augmented model (A), are much superior to those constructed under model (A). The bias is reduced for areas, large or small.

**Table 4**  
Monte Carlo Properties of individual area predictors using versions of model (A) (10,000 samples generated by model B)

State	Quantity	EBLUP	PB	ITF	YR	AUG2
1	Bias	-0.597	-0.225	-0.052	-0.473	-0.030
	MSE	1.471	1.165	1.130	1.331	1.139
2	Bias	-0.496	-0.227	-0.170	-0.358	-0.070
	MSE	1.330	1.134	1.115	1.207	1.124
14	Bias	-0.121	0.004	-0.031	0.055	-0.025
	MSE	1.100	1.085	1.086	1.089	1.105
26	Bias	0.057	0.157	0.115	0.249	0.053
	MSE	1.126	1.148	1.136	1.188	1.132
38	Bias	0.380	0.453	0.406	0.601	0.202
	MSE	1.340	1.405	1.361	1.571	1.233
50	Bias	0.922	0.980	0.931	1.178	0.537
	MSE	2.196	2.316	2.215	2.767	1.577

## 6. Conclusions

In this paper, several calibrated predictors are reviewed. We offer a fresh look at the benchmarking restriction (7). Imposing the restriction is viewed as an adjustment problem and a criterion that unifies the derivation of calibrated predictors is presented. The criterion approach to the problem opens the door for consideration of other predictors.

Implicit in the imposition of the restriction is the possibility that the small area model is misspecified and the predictors are biased. When the model is misspecified, the calibration adjustment only adjusts for the overall bias, not for the bias at the small area level. The augmented model approach leads to a self-calibrated predictor and reduces the bias at the small area level. Also, variance estimation for the externally calibrated predictors is relatively complex, while variance estimation for self-calibrated predictors is straightforward. In summary, if the bias is a concern, use of a self-calibrated augmented model is preferred to external calibration.

## 7. Appendix

**Proof of Theorem 1:** Let  $\hat{y}_i^H$  be the BLUP of  $y_i$  and let  $\hat{y}_i$  be any linear unbiased predictor of  $y_i$ . By standard results for BLUP (see, for example, Robinson (1991) and Harville (1976)), we have

$$\text{cov}(\hat{y}_i^H - y_i, \hat{y}_j - \hat{y}_j^H) = 0, \quad \text{if } i \neq j, \quad (\text{A.1})$$

for  $i = 1, \dots, n$  and  $j = 1, \dots, n$ . Let  $R(\hat{\mathbf{y}}^a)$  denote the collection of all linear unbiased predictors that satisfy (7). For any  $\hat{\mathbf{y}} \in R(\hat{\mathbf{y}}^a)$ , by (A.1), we have

$$E\{(\hat{y}_i - y_i)^2\} = E\{(\tilde{y}_i^H - y_i)^2\} + E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \quad (\text{A.2})$$

Therefore,

$$\begin{aligned} Q(\hat{y}) &= \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - y_i)^2\} \\ &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - \tilde{y}_i^H)^2\}. \quad (\text{A.3}) \end{aligned}$$

Since  $\hat{y}$  satisfies (7), we have  $\sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i$  and, for the  $\hat{y}_i^{\bar{a}}$  defined in (19),

$$\begin{aligned} \hat{y}_i^{\bar{a}} &= \tilde{y}_i^H + \tilde{a}_i \left[ \sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right] \\ &= \tilde{y}_i^H + \tilde{a}_i \left[ \sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]. \end{aligned}$$

By (A.1),

$$\begin{aligned} Q(\hat{y}_i^{\bar{a}}) &= \sum_{i=1}^n \varphi_i E(\hat{y}_i^{\bar{a}} - y_i)^2 \\ &= \sum_{i=1}^n \varphi_i E \left\{ \left[ (\tilde{y}_i^H - y_i) + \tilde{a}_i \left( \sum_{j=1}^n \omega_j \hat{y}_j - \sum_{j=1}^n \omega_j \tilde{y}_j^H \right) \right]^2 \right\} \\ &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} + E \left\{ \left[ \sum_{j=1}^n \omega_j (Y_j - \tilde{y}_j^H) \right]^2 \right\} \sum_{i=1}^n \varphi_i \tilde{a}_i^2. \quad (\text{A.4}) \end{aligned}$$

Since  $\tilde{a}_i = (\sum_{i=1}^n \varphi_i^{-1} \omega_i^2)^{-1} \varphi_i^{-1} \omega_i$  in (A.4), we have

$$\begin{aligned} Q(\hat{y}_i^{\bar{a}}) &= \sum_{i=1}^n \varphi_i E\{(\tilde{y}_i^H - y_i)^2\} \\ &\quad + E \left\{ \left[ \sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} \left( \sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right)^{-1}. \quad (\text{A.5}) \end{aligned}$$

Note that

$$\begin{aligned} E \left\{ \left[ \sum_{j=1}^n \omega_j (\hat{y}_j - \tilde{y}_j^H) \right]^2 \right\} &\leq \sum_{j=1}^n \sum_{k=1}^n \omega_j \omega_k g_j g_k \\ &= \left( \sum_{i=1}^n \omega_i g_i \right)^2, \quad (\text{A.6}) \end{aligned}$$

where  $g_j = \{E[(\hat{y}_j - \tilde{y}_j^H)^2]\}^{0.5}$ . By Cauchy's inequality,

$$\begin{aligned} \left( \sum_{j=1}^n \omega_j g_j \right)^2 &\leq \left( \sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left( \sum_{i=1}^n \varphi_i g_i^2 \right) \\ &= \left( \sum_{i=1}^n \varphi_i^{-1} \omega_i^2 \right) \left[ \sum_{i=1}^n \varphi_i E\{(\hat{y}_i - \tilde{y}_i^H)^2\} \right]. \quad (\text{A.7}) \end{aligned}$$

Combining (A.3), (A.5), (A.6), and (A.7), we have  $Q(\hat{y}_i^{\bar{a}}) \leq Q(\hat{y})$ .

To show the uniqueness of  $\hat{y}_i^{\bar{a}}$ , we need to check when the inequalities (A.6) and (A.7) become equalities. Inequality (A.6) becomes an equality if and only if

$$\hat{y}_j - \tilde{y}_j^H = c_j^0 + c_j^1 (\hat{y}_1 - \tilde{y}_1^H) \quad (\text{A.8})$$

for some constants  $c_j^0$  and  $c_j^1$ ,  $j = 2, \dots, n$ . Inequality (A.7) becomes an equality if and only if

$$\sqrt{\varphi_1^{-1} \omega_1^2} \sqrt{v_j g_j^2} - \sqrt{v_j^{-1} \omega_j^2} \sqrt{\varphi_1 g_1^2} = 0, \quad (\text{A.9})$$

or, equivalently,

$$\varphi_1^2 \omega_1^{-2} E\{(\hat{y}_1 - \tilde{y}_1^H)^2\} = v_j^2 \omega_j^{-2} E\{(\hat{y}_j - \tilde{y}_j^H)^2\}. \quad (\text{A.10})$$

Also,

$$\sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i. \quad (\text{A.11})$$

Combining (A.8), (A.10), and (A.11), we have that the equality holds if and only if  $\hat{y}_j = \hat{y}_j^{\bar{a}}$ . Thus, we have shown that  $\hat{y}_i^{\bar{a}}$  is the unique linear unbiased predictor that satisfies (7) and minimizes criterion (18).

## Acknowledgements

This research was funded in part by cooperative agreement 68-3A75-14 between the USDA Natural Resources Conservation Service and Iowa State University.

## References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of American Statistical Association*, 28-36.
- Harville, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *Annals of Statistics*, 384-395.
- Henderson, C.R. (1963). Selection index and expected genetic advance. In *Statistical Genetics and Plant Breeding*, (Eds., W.D. Hanson and H.F. Robinson), National Academy of Sciences-National Research Council, Washington, Publication 982, 141-163.
- Isaki, C.T., Tsay, J.H. and Fuller, W.A. (2000). Estimation of census adjustment factors. *Survey Methodology*, 31-42.
- Mantel, H.J., Singh, A.C. and Barreau, M. (1993). Benchmarking of small area estimators. In *Proceedings of International Conference on Establishment Surveys*, American Statistical Association, Washington, DC, 920-925.
- Pfeffermann, D., and Barnard, C.H. (1991). Some new estimators for small-area means with application to the assessment of Farmland values. *Journal of Business & Economic Statistics*, 73-84.



- Rao, J.N.K. (2003). *Small area estimation*. New York: John Wiley & Sons, Inc.
- Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, 15-51.
- Wang, J., and Fuller, W.A. (2003). The mean square error of small area estimators constructed with estimated area variances. *Journal of American Statistical Association*, 716-723.
- You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 431-439.
- You, Y., and Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 197-208.

# A Bayesian allocation of undecided voters

Balagobin Nandram and Jai Won Choi<sup>1</sup>

## Abstract

Data from election polls in the US are typically presented in two-way categorical tables, and there are many polls before the actual election in November. For example, in the Buckeye State Poll in 1998 for governor there are three polls, January, April and October; the first category represents the candidates (*e.g.*, Fisher, Taft and other) and the second category represents the current status of the voters (likely to vote and not likely to vote for governor of Ohio). There is a substantial number of undecided voters for one or both categories in all three polls, and we use a Bayesian method to allocate the undecided voters to the three candidates. This method permits modeling different patterns of missingness under ignorable and nonignorable assumptions, and a multinomial-Dirichlet model is used to estimate the cell probabilities which can help to predict the winner. We propose a time-dependent nonignorable nonresponse model for the three tables. Here, a nonignorable nonresponse model is centered on an ignorable nonresponse model to induce some flexibility and uncertainty about ignorability or nonignorability. As competitors we also consider two other models, an ignorable and a nonignorable nonresponse model. These latter two models assume a common stochastic process to borrow strength over time. Markov chain Monte Carlo methods are used to fit the models. We also construct a parameter that can potentially be used to predict the winner among the candidates in the November election.

**Key Words:** Markov chain Monte Carlo; Metropolis sampler; Multinomial-Dirichlet model; Time-dependent model; Two-way categorical table.

## 1. Introduction

It is a common practice to use two-way categorical tables to present survey data. Our application is to predict the winner in an election using tables constructed from a short series of polls taken before the actual election. For many surveys, there are missing data and this gives rise to partial classification of the sampled individuals. Little and Rubin (2002, section 1.3) give definitions of the three missing data mechanism (missing completely at random - MCAR, missing at random - MAR, missing not at random - MNAR); ignorable models are used to analyze data from MAR and MCAR mechanisms and nonignorable models for data from MNAR mechanisms. Thus, for the two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing). One may not know how the data are missing, and a model that includes some difference between the observed data and missing data (*i.e.*, nonignorable missing data) may be preferred. For a general  $r \times c$  categorical table we address the issue of estimation of the cell probabilities of the two-way table. This problem is important because, with a substantial number of undecided voters, an election prediction based on only the partially observed data may be misleading.

As in Nandram, Cox and Choi (2005) essentially there are four two-way tables, one table with all complete cases and three supplemental tables. Of the three supplemental tables, the first has only row classification (item

nonresponse), the second has only column classification (item nonresponse), and the third does not have any classification (unit nonresponse). We have extended the ignorable and nonignorable nonresponse models for two-way categorical tables of Nandram, *et al.* (2005) to accommodate a third category (*i.e.*, time in a short sequence of election polls). We have extended these models even further to include a time-dependent nonignorable nonresponse structure. The inclusion of the time-dependent structure can provide a more efficient prediction. A Bayesian method permits modeling different patterns of missingness under the ignorability and nonignorability assumptions, and a time-dependent nonignorable nonresponse model is obtained.

Our application is in Ohio governor's election, and there are several related problems. The sampled persons are categorized by two types of attributes and the cells of such categorical tables are analyzed. However, only partial classification of the individuals is available because some individuals are classified by at most one attribute, and others are left unclassified. Specifically, we use tabular data from the Ohio polls to study the relation between a measure of voters' status (likely to vote and unlikely to vote) and candidate preference (Fisher, Taft and other) to illustrate our methodology. It is interesting that voters' status is related to candidate preference. Also, it is desirable to make an adjustment for undecided voters because the proportion of undecided voters is usually high, and they often decide the final outcome of an election.

1. Balagobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institute, 100 Institute Road, Worcester MA 01609. E-mail: balnan@wpi.edu.; Jai Won Choi, Department of Biostatistics, Medical College of Georgia, 1420 15<sup>th</sup> Street, Augusta, GA 30912. E-mail: jchoi@MCG.edu.

We do not know whether an ignorable nonresponse model or a nonignorable nonresponse model is appropriate, but one may have uncertainty about the ignorability of undecided voters in election polls. Referring to the Buckeye State Poll, Chen and Stasny (2003) stated that "The assumption of nonignorability of the nonresponse may be a reasonable assumption in this study because people might be reluctant to express their preference for an unpopular candidate, or if their current preferences are not firm or accurate enough for the standards of the interview." They also said that while Chang and Krosnick (2001) use ignorable models for their analyses, Chang and Krosnick (2001) suggested that nonresponse might be related to the unobserved data itself. Chen and Stasny (2003) fit three ignorable nonresponse models (A, B and C) and one nonignorable nonresponse model (D). We compare our results with theirs.

Nandram and Choi (2002 a, b) use an expansion model to study nonignorable nonresponse binary data. The expansion model, a nonignorable nonresponse model, degenerates into an ignorable nonresponse model (in the spirit of Draper 1995). This degeneracy occurs when a parameter in the nonignorable nonresponse model is set to a certain value; a good description of the centering idea is given in Nandram, *et al.* (2005, section 1.2). Because it is difficult to carry out this procedure as described, we use an alternative procedure as in Nandram, *et al.* (2005). This permits an expression of uncertainty about ignorability. This is the idea of centering a nonignorable nonresponse model on an ignorable nonresponse model, and we have used it in several of our papers to express uncertainty about ignorability or nonignorability. Here, for nonignorable nonresponse we attempt a related methodology, but the issues for a two-way categorical table are more complex, especially when a third category (*i.e.*, time) is included in these tables.

Using the approach of Chen and Fienberg (1974), Chen and Stasny (2003) describe the two issues we are discussing in this paper. For the two-way categorical tables they can handle item nonresponse only; unit nonresponse is excluded from their analysis. However, they assume that the data are missing at random and show how to obtain maximum likelihood estimators under their model. They also use a nonignorable nonresponse model (D), which they claim is their best model. It is noted in Little and Rubin (2002, chapter 15) that one issue of the nonignorable nonresponse model for this problem is that there are too many parameters, and many parameters are not identified, so they attempted a correction using hierarchical log-linear models. See Nandram, *et al.* (2005) for the case in which there are three supplemental tables.

Our methodology differs from those of Chen and Stasny (2003). The major difference is that we use a Bayesian approach. This permits us to use a method that does not rely on asymptotic theory, incorporate nonignorable missingness into the modeling and obtain time-dependent nonignorable model for estimating the proportion of voters for the three candidates. Looking to predict the winner more convincingly, we have also constructed a new parameter; it is relatively easy to analyze this parameter within the Bayesian paradigm. The Bayesian method permits modeling different patterns of missingness under two different assumptions (*i.e.*, ignorable and nonignorable missingness). Our idea is to start with an ignorable nonresponse model, which is then expanded into a nonignorable nonresponse model, and to the time-dependent nonignorable nonresponse model. It is worth noting that unit nonresponse is also included in our modeling which the other researchers consider as a separate problem using weighting adjustment (*e.g.*, see discussion in Kalton and Kasprzyk 1986). However, there can be nonignorability here as well, and one would need to include unit and item nonresponses simultaneously.

In this paper, our key contribution is to introduce a Bayesian method to analyze data from an  $r \times c$  categorical table when there are both item and unit nonresponse, and the missing data mechanism can be nonignorable with a time-dependent structure. In Section 2, we describe the categorical data on voters' status and candidate preference with a time-dependent structure. In Section 3, we describe the methodology to obtain estimates of the cell probabilities incorporating the two types of missing data, and we show how to expand an ignorable nonresponse model into a nonignorable nonresponse model and time-dependent model. We also show how to use Markov chain Monte Carlo methods to fit the nonignorable nonresponse model. In Section 4, we analyze the Ohio election data to demonstrate the versatility of our methods. Finally, Section 5 has concluding remarks.

## 2. Data on 1998 Ohio Polls

The Center for Survey Research (CSR) at the Ohio State University conducted the Buckeye State Poll (BSP) during the 1998 election for Senator, Governor, Attorney General, State Secretary, Treasurer and Columbus Mayor. In certain months before the election, CSR conducted pre-election surveys as part of the BSP and included additional questions to collect information related to the respondent's likelihood of voting and candidate preference. In the BSP, households are sampled using the Random Digit Dialing (RDD) method, and one adult per household is selected to be interviewed using the last birthday method (Lavrakas 1993).



It is pertinent to briefly describe the RDD method. Polling firms make extensive use RDD, and the main goal of RDD is to develop a representative sample of the overall voter population. RDD sampling assumes that a representative sample cannot be obtained using listed telephone numbers in the directory. Each telephone number has 10 digits, the first three form the area code, the next three form the prefix (colloquially called the exchange), and the last four (suffix) identify a particular subscriber or a household (one household can have more than one phone number). The area codes are geographically based and typically identify localities in a state, and the exchanges can be geographically oriented. There are ten million numbers to dial but roughly less than 25% of these are real telephone numbers. Thus time and money are wasted in dialing unused numbers. We discuss this further in Section 3.

Chen and Stasny (2003) and Chang and Krosnick (2001) analyzed data from three BSP pre-election forecasting polls. Details of each of these three BSP pre-election surveys can be found in Table 1. These BSP pre-election surveys measured respondents' candidate preferences three times (January, April and October) for the November 1998 Ohio Governor race. In addition, respondents were asked for their self-reported likelihood of voting in the upcoming election using two questions. Chang and Krosnick (2001) also used filter variables (such as registered to vote, self-reported likelihood of voting, and voted in the last major election, *etc.*) to obtain those most likely to vote. Thus prediction is based only on the respondents likely to vote. Those registered to vote are classified into likely to vote, unlikely to vote and undecided. Chang and Krosnick (2001) showed that deterministic allocation of undecided respondents provide improvement in forecasting voters' candidate preferences, as compared to exclusion of all undecided respondents. Chen and Stasny (2003) used probability models to allocate the undecided voters and compared their forecasting with that of Chang and Krosnick (2001).

The data set in Chen and Stasny (2003) is slightly different because we use the undecided counts (unit nonresponse) on both variables. A voter can be undecided on at least one of the two categorical variables at each of the three polls. Chen and Stasny (2003) only study the data with undecided in exactly one variable, not both. In Table 1 for the undecided voters in both variables the counts for the January, April and October polls are respectively 5, 3 and 4; these numbers are bolded. In fact, the inclusion of these counts into our model, is an extension of the models in Chen and Stasny, and generalizes our methodology considerably.

We briefly describe the 2 × 3 categorical table of Ohio election data by voters' status (VS) and candidate preference (CAN). Here VS is a binary variable, and there are two levels: likely to vote and not likely to vote; CAN has

three levels: Fisher, Taft, others. There are also undecided voters in VS and CAN. The bulk of the undecided voters come from voters who are "likely to vote" and "unlikely to vote" and the numbers are 173, 142 and 138 for January, April and October respectively; the undecided voters for Fisher, Taft and others are much smaller.

**Table 1**  
**Classification of October 1998 Buckeye State Poll by voting status and candidate**

Status	Candidate				Total
	Fisher	Taft	Other	Undecided	
a. January, 1998					
Likely to vote	127	183	8	109	427
Not likely to vote	57	94	4	59	214
Undecided	0	2	0	5	7
Total	184	279	12	173	648
b. April, 1998					
Likely to vote	114	135	1	61	311
Not likely to vote	104	149	3	78	334
Undecided	2	6	0	3	11
Total	220	290	4	142	656
c. October, 1998					
Likely to vote	112	140	23	61	336
Not likely to vote	96	108	21	73	298
Undecided	7	11	1	4	23
Total	215	259	45	138	657

NOTE: These data are taken from Chang and Krosnick (2001); Chen and Stasny (2003) used a very similar data set; they did not use 5, 3, 4, the number of undecided voters in both variables.

In the January 1998 poll, about 73% of the voters are completely classified, 27% have no decision about candidate preference, only 1% did not know whether they would vote or not, and only five persons were completely unclassified among the 648 participants. The data set, used in our study, is presented in Table 1 as a 2 × 3 categorical table of voters' status and candidate preference. Our problem is to predict the winning candidate by estimating the proportion of final votes for each candidate.

The samples obtained in January, April and October are independent. There is no oversampling for a particular sub-population or weighting of the original sample. Like many telephone surveys, RDD frame suffers from the common problem of undercoverage. As telephone coverage is not uniform over age, race, sex, income and geography, there is a need to poststratify the original sample to reduce the coverage bias by properly weighting the original data.

We perform a preliminary test of heterogeneity of the cell proportions across the three polls. Assuming a missing at random mechanism, we fill in the undecided votes. We assume that for each row (column) the undecided voters are filled in proportionally to the cell counts. Let  $n_{ijk}$  denote the adjusted cell counts with  $n_i = \sum_{j=1}^r \sum_{k=1}^c n_{ijk}$ , and let  $p_{ijk}$  denote the cell proportions. For a model of heterogeneous proportions, we assume that

$$n_i | p_i \sim \text{Multinomial}(n_i, p_i) \text{ and } p_i \sim \text{Dirichlet}(\mathbf{1}), i=1, \dots, T,$$

where  $\mathbf{1}$  is a  $rc$ -vector of ones.

For a model of homogeneous proportions, we assume that

$$\mathbf{n}_t | \mathbf{p} \sim \text{Multinomial}(n_t, \mathbf{p}), t=1, \dots, T, \text{ and } \mathbf{p} \sim \text{Dirichlet}(\mathbf{1}).$$

Then, the Bayes factor of heterogeneity versus homogeneity is

$$\text{BF} = \frac{1}{\{(rc-1)!\}^2} \left[ \prod_{j=1}^r \prod_{k=1}^c \left\{ \frac{\prod_{t=1}^T n_{tjk}!}{\left( \sum_{t=1}^T n_{tjk} \right)!} \right\} \right] \frac{\left\{ \sum_{t=1}^T n_t + rc - 1 \right\}!}{\prod_{t=1}^T (n_t + rc - 1)!}.$$

Thus, using the adjusted cell counts, the logarithm of the Bayes factor (LBF) is approximately 12.4, showing very strong evidence for heterogeneity, and supporting our time-dependent model.

In a similar manner, we have computed the Bayes factors of  $p_1 = p_2 \neq p_3$  or  $p_1 \neq p_2 = p_3$  versus homogeneity; the LBFs are 7.6 and 4.4 respectively. Thus, the time-dependence occurs for both periods, January-April and April-October.

### 3. Methodology

We have constructed a time-dependent nonignorable nonresponse model for the 1998 Ohio Poll data. For comparison we have also considered two other models, an ignorable and a nonignorable nonresponse model. These latter two models are not time-dependent because we assume that the three time points come from the same stochastic process (*i.e.*, no correlation across time). Our main contribution is the time-dependent model. We have used the ignorable and nonignorable nonresponse models for a single time point in Nandram, *et al.* (2005). Although these two models are not appropriate in the present context, they are natural to motivate our time dependent nonignorable nonresponse model. Essentially we start with the ignorable nonresponse model which is expanded into a nonignorable nonresponse model, and we extend the nonignorable nonresponse model to a time-dependent model.

In RDD stratification and clustering are used to reduce the excess artificial numbers. Stratification by area code and some exchanges is used; geographic ordering (state or region) with systematic selection provides implicit stratification of exchanges. If an exchange is used to form a stratum, there are still ten thousand numbers to dial, still a large waste with numerous redundant numbers. The Mitofsky-Waksberg (see Waksberg 1978) procedure is a stratified two-stage cluster sampling design used to reduce the artificial numbers. Exchange areas are divided into equal size, and a random sample of exchanges is taken with

replacement from those eligible (according to the measure of size of each exchange area). Within selected exchange area, a fixed number of telephone numbers is generated at random, without replacement and dialed. Thus, there is also differential probabilities of selection (*i.e.*, unequal cluster sizes) that must be considered in a comprehensive analysis. There are other variants of this procedure. RDD was adequate in 1998 Ohio election, but because of new technological innovations (*e.g.*, cellular phone, email, internet, *etc.*), the usefulness of RDD may be diminished. In this paper, our method and models do not include stratification, clustering or differential probabilities of selection.

Our models are used to estimate the proportions of voters voting for Fisher, Taft and other in the October poll. Then, assuming no catastrophic change in the November election, we predict the proportion of voters voting for Fisher, Taft and other. In this way we can predict the winner in the November election. We are excited by a referee's suggestion that one can use a mixture model to cover the possibility of a catastrophe.

In Sections 3.1 and 3.2 we describe the notations and the three models. In Section 3.3 we show how to fit the time-dependent nonignorable nonresponse model. The ignorable and nonignorable nonresponse models can be fit in a similar manner (see Nandram, *et al.* 2005 for details). In Section 3.4 we show how to specify the two parameters ( $\mu_0$  and  $c_0^2$ ), and in Section 3.5 we show how to do estimation in the October poll and prediction in the November election.

#### 3.1 Notation

Let  $I_{jkt} = 1$  if the  $\ell^{\text{th}}$  voter belongs to the  $j^{\text{th}}$  row and  $k^{\text{th}}$  column of the two-way table at time  $t$  and  $I_{jkt} = 0$  otherwise,  $t = 1, \dots, T$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ ,  $\ell = 1, \dots, L$ . That is,  $I_{jkt} = 1$  denotes the cell of the  $r \times c$  table that a voter belongs to. In our application  $T = 3$ ,  $r = 2$  and  $c = 3$ . Let  $J_{st} = 1$  if the  $\ell^{\text{th}}$  voter falls in table  $s$  ( $s = 1, 2, 3, 4$ ) and  $J_{st} = 0$  otherwise,  $s = 1, \dots, 4$ ,  $\sum_{s=1}^4 J_{st} = 1$ ;  $J_{st}$  indicates which table an individual belongs to and  $\mathbf{J}_{t\ell} = (J_{1t\ell}, J_{2t\ell}, J_{3t\ell}, J_{4t\ell})$ .

Let the cell counts be  $y_{tjk} = \sum_{\ell=1}^n I_{tjk\ell} J_{st\ell}$ ,  $s = 1, 2, 3, 4$  for the four tables at each poll. Here  $y_{1jk}$  are observed and  $y_{2jk}$ ,  $s = 2, 3, 4$ ,  $t = 1, \dots, T$  are missing (*i.e.*, latent variables). For  $y_{1jk}$  we know that  $\sum_{j=1}^r \sum_{k=1}^c y_{1jk} = n_{t0}$ , the number of individuals with complete data. For  $y_{2jk}$  we know that  $\sum_{k=1}^c y_{2jk} = u_{tj}$ , where the row margins  $u_{tj}$ ,  $j = 1, \dots, r$  are observed. For  $y_{3jk}$  we know that  $\sum_{j=1}^r y_{3jk} = v_{tk}$ , where the column margins  $v_{tk}$ ,  $k = 1, \dots, c$  are observed. For  $y_{4jk}$  we know that  $\sum_{j=1}^r \sum_{k=1}^c y_{4jk} = w_t$  (unit nonresponse). In this analysis  $n_{t0}$ ,  $u_{tj}$ ,  $v_{tk}$  and  $w_t$  are held fixed (*i.e.*, fixed margin analysis) and known.

Whenever it is convenient, we will use notations such as

$$\sum_{s,j,k} y_{tsjk} \equiv \sum_{s=1}^4 \sum_{j=1}^r \sum_{k=1}^c y_{tsjk}, \quad \prod_{s,j,k} \pi_{tsjk} \equiv \prod_{s=1}^4 \prod_{j=1}^r \prod_{k=1}^c \pi_{tsjk}$$

and  $y_{t(1)} = (y_{t2}, y_{t3}, y_{t4})$ ,  $y_{t(2)} = (y_{t1}, y_{t3}, y_{t4})$ , etc., where  $y_{ts} = (y_{tsjk}, j = 1, \dots, r, k = 1, \dots, c, t = 1, \dots, T, s = 1, 2, 3, 4)$ . Also, we let  $y_1 = (y_{11}, \dots, y_{T1})$  and  $y_{(1)} = (y_{1(1)}, \dots, y_{T(1)})$  with  $y_{(1)} = (y_{t(1)}, \dots, y_{t(4)})$ . Also,  $\sum_{s,j,k}^{4,r,c} y_{tsjk} = n_t$ . We will also use  $y_{ts-} = \sum_{j,k} y_{tsjk}$ ,  $y_{t-jk} = \sum_s y_{tsjk}$ , etc.,  $y_t = (y_{t1}, y_{t2}, y_{t3}, y_{t4})$  and  $y = (y_1, \dots, y_T)$ .

### 3.2 Nonresponse models

Letting  $I_{t\ell} = (I_{tj\ell}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c, \ell = 1, \dots, L)$ , for all models, we take

$$I_{t\ell} | p_t \sim \text{Multinomial}\{1, p_t\}, \quad (1)$$

where

$$\sum_{j=1}^r \sum_{k=1}^c p_{tjk} = 1, p_{tjk} \geq 0, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c.$$

For the ignorable nonresponse model we take

$$J_{t\ell} | \pi_t \sim \text{Multinomial}\{1, \pi_t\}. \quad (2)$$

That is, there is no dependence on the cell status of an individual. For the nonignorable nonresponse models we take

$$J_{t\ell} | \{I_{tj\ell} = 1, I_{tj'k'\ell} = 0, j \neq j', k \neq k', \pi_{tjk}\} \sim \text{Multinomial}\{1, \pi_{tjk}\}. \quad (3)$$

Assumption (3) specifies that the probabilities an individual belongs to one of the four tables depend on the two characteristics (*i.e.*, row and column classifications) of the individual. In this manner we incorporate the assumption that the missing data is nonignorable. Note that conditional on the specified parameters in (1)-(3), one voter's behavior is correlated with another at the same time  $t$ , but there is independence over time. It is worth noting here that while the parameters in (2) are identifiable, those in (3) are not identifiable. This is where the difficulty in the nonignorable nonresponse model arises, and special attention is needed.

It follows from (1) and (2) that for the ignorable model

$$g(p, \pi | y) \propto \prod_{t=1}^T \left[ \prod_{s=1}^4 \pi_{ts-} \left[ \prod_{j=1}^r \prod_{k=1}^c \frac{p_{tjk}^{y_{tsjk}}}{y_{tsjk}!} \right] \right] \quad (4)$$

subject to  $\sum_{k=1}^c y_{t2jk} = u_{tj}$ ,  $j = 1, \dots, r$ ,  $\sum_{j=1}^r y_{t3jk} = v_{tk}$ ,  $k = 1, \dots, c$ , and  $\sum_{j=1}^r \sum_{k=1}^c y_{t4jk} = w_t$ . Note that under ignorability the likelihood function in (4) separates into two pieces, one that contains the  $\pi_{ts}$  only and the other the  $p_{tjk}$ ,

and inference about these two parameters are unrelated; see Section 3.2 of Nandram, *et al.* (2005) for the original discussion of this model. Also, it follows from (1) and (3) that for the nonignorable nonresponse models the augmented likelihood function for  $p, \pi, y_{(1)} | y_1$  is

$$g(p, \pi, y_{(1)} | y_1) \propto \prod_{t=1}^T \left[ \prod_{s,j,k}^{4,r,c} \frac{\pi_{tsjk}^{y_{tsjk}}}{y_{tsjk}!} \left[ \prod_{j,k}^{r,c} p_{tjk}^{y_{t,jk}} \right] \right] \quad (5)$$

subject to  $\sum_{k=1}^c y_{t2jk} = u_{tj}$ ,  $j = 1, \dots, r$ ,  $\sum_{j=1}^r y_{t3jk} = v_{tk}$ ,  $k = 1, \dots, c$ , and  $\sum_{j=1}^r \sum_{k=1}^c y_{t4jk} = w_t$ ; see Nandram, *et al.* (2005) for a description of identifiability in a similar situation.

For the ignorable and nonignorable nonresponse models, we take

$$p_t | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2, \tau_2), t = 1, \dots, T+1, \quad (6)$$

where we consider prediction at  $T+1$ , one step ahead (November). The probabilistic structure in (6) permits a “borrowing of strength” across time. Note that the  $k$ -dimensional vector  $x$  has a Dirichlet distribution if  $p(x|\alpha) = \prod_{j=1}^k x_j^{\alpha_j-1} / D(\alpha)$ ,  $x_j \geq 0$ ,  $j = 1, \dots, k$ ,  $\sum_{j=1}^k x_j = 1$ , where  $D(\alpha)$  is the Dirichlet function and  $\alpha_j > 0$ ,  $j = 1, \dots, k$ . For a quick reference see Ghosh and Meeden (1997, pages 42, 50, 127) in connection with the Polya urn distribution, and more appropriately its use as a conjugate prior in multinomial sampling; starting with our first paper (*i.e.*, Nandram 1998) we have been using the Dirichlet-multinomial extensively in our research.

We next describe the stochastic models for the  $\pi_{tjk}$ . For the ignorable nonresponse model, we take

$$\pi_t \sim \text{Dirichlet}(\mathbf{1}), t = 1, \dots, T, \quad (7)$$

where  $\mathbf{1}$  is a four-dimensional vector of ones. We need (7) because  $T$  is small (*i.e.*,  $T = 3$  in our application). Thus, we use the uniform prior in  $R^4$  (essentially noninformative); otherwise we will have to specify the unknown parameters of the Dirichlet distribution with virtually no data. For the nonignorable nonresponse models we take

$$\pi_{tjk} | \mu_1, \tau_1 \sim \text{Dirichlet}(\mu_1, \tau_1), t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c. \quad (8)$$

First, we note that (8) provides a “borrowing of strength” across time. More importantly, because  $\pi_{tjk}$  are not identifiable so are  $\mu_1$  and  $\tau_1$ . One possible way out of this dilemma is to “center” the nonignorable nonresponse model on the ignorable nonresponse model.

For the time-dependent model, we take

$$p_t | p_{t-1}, \tau_2 \sim \text{Dirichlet}(p_{t-1}, \tau_2), t = 1, \dots, T+1, \quad (9)$$



where  $\mathbf{p}_0$  is also unknown. Note that

$$E\{\mathbf{p}_t \mid \mathbf{p}_{t-1}, \tau_2\} = \mathbf{p}_{t-1}, t = 1, \dots, T+1;$$

so that  $\{\mathbf{p}_t\}$ , *a priori*, is a martingale vector. Here  $T$  is small (*i.e.*,  $T = 3$ ). Thus, this time-dependent structure seems more appropriate, and can potentially provide improved precision. Note also that we have taken  $\mathbf{p}_0 \sim \text{Dirichlet}(\mathbf{1})$ .

Finally, we specify prior densities for the hyperparameters. First, we take

$$\mu_1, \mu_2 \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{1}), \quad (10)$$

essentially noninformative prior densities.

Finally,  $\tau_1$  and  $\tau_2$  are independent and identically distributed random variables from

$$f(x) = 1/(1+x)^2, \quad x \geq 0. \quad (11)$$

Again this is an essentially noninformative prior density. Note that  $\mu_1$  and  $\tau_1$  do not exist in the ignorable nonresponse model. Gelman (2006) recommended priors like (11) instead of the ill-behaved proper diffuse gamma priors.

For the nonignorable nonresponse models we need to be more careful to specify the prior density of  $\tau_1$  because  $\pi_{ijk}$  are not identifiable. Here we attempt to “center” the nonignorable nonresponse models on the ignorable nonresponse model. In (8) the parameter  $\tau_1$  tells us about the closeness of the nonignorable model to the ignorable model. For example, if  $\tau_1$  is small, the  $\pi_{ijk}$  will be very different, and if  $\tau_1$  is large, the  $\pi_{ijk}$  will be very similar. Thus, *a priori* inference will be sensitive to the choice of  $\tau_1$ , and one has to be careful in choosing  $\tau_1$ . We would like to choose a prior density for  $\tau_1$  so that the nonignorable nonresponse model is kept close to the ignorable nonresponse model. Thus, we take

$$\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2), \quad (12)$$

where  $E(\tau_1) = \mu_0$  and  $\text{CV}(\tau_1) = c_0$ , with  $\text{CV}$  the coefficient of variation; both  $\mu_0$  and  $c_0$  are to be specified. We use the prior (12) because by an appropriate choice of  $\mu_0$  and  $c_0$  it is possible to center the nonignorable nonresponse model on the ignorable nonresponse model. Of course, one can use other convenient proper priors with parameters like  $\mu_0$  and  $c_0$  to facilitate the centering. In Section 3.4 we will use samples from the posterior density of  $\tau_1$  under the ignorable nonresponse model to specify  $\mu_0$  and  $c_0$ .

For each of the three models, it is easy to write down the joint prior density of the parameters. For example, for the time-dependent model the joint prior density is

$$p(\mathbf{p}, \boldsymbol{\pi}, \mu_1, \tau_1, \tau_2) \propto \tau_1^{1/c_0^2-1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2} \\ \times \prod_{t=1}^T \left\{ \frac{\prod_{j=1}^r \prod_{k=1}^c p_{tjk}^{p_{t-1,jk}\tau_2-1}}{D(\mathbf{p}_{t-1}\tau_2)} \prod_{j=1}^r \prod_{k=1}^c \frac{\prod_{s=1}^4 \pi_{tsjk}^{\mu_{1s}\tau_1-1}}{D(\mu_1\tau_1)} \right\}, \quad (13)$$

where  $D(\cdot)$  is the Dirichlet function.

### 3.3 Fitting the time-dependent nonignorable nonresponse model

Combining the likelihood function in (5) with the joint prior density in (13) via Bayes' theorem, the joint posterior density of the parameters  $\boldsymbol{\pi}, \mathbf{p}, \mu_1, \tau_1, \tau_2$  and the latent variables  $\mathbf{y}_{(1)}$  is

$$\pi(\mathbf{p}, \boldsymbol{\pi}, \mu_1, \tau_1, \tau_2, \mathbf{y}_{(1)} \mid \mathbf{y}_{(1)}) \\ \propto \tau_1^{1/c_0^2-1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1+\tau_2)^2} \prod_{t=1}^T \left[ \prod_{j,k}^{4,r,c} \frac{\pi_{tsjk}^{y_{tsjk}}}{\pi_{tsjk}^1} \prod_{j,k}^{r,c} p_{tjk}^{y_{tjk}} \right] \\ \times \prod_{t=1}^T \left\{ \frac{\prod_{j=1}^r \prod_{k=1}^c p_{t-1,jk}^{p_{t-1,jk}\tau_2-1}}{D(\mathbf{p}_{t-1}\tau_2)} \prod_{j=1}^r \prod_{k=1}^c \frac{\prod_{s=1}^4 \pi_{tsjk}^{\mu_{1s}\tau_1-1}}{D(\mu_1\tau_1)} \right\} \quad (14)$$

subject to  $\sum_{k=1}^c y_{t2,jk} = u_{tj}$ ,  $j = 1, \dots, r$ ,  $\sum_{j=1}^r y_{t3,jk} = v_{tk}$ ,  $k = 1, \dots, c$ , and  $\sum_{j=1}^r \sum_{k=1}^c y_{t4,jk} = w_t$ ,  $t = 1, \dots, T$ .

The posterior density in (14) is complex, so we will use Markov chain Monte Carlo methods to fit it. However, it is easy to fit the time-dependent model using the griddy Metropolis-Hastings sampler (our terminology) as we will describe. Also, in a similar manner using the griddy Gibbs sampler (Ritter and Tanner 1992), it is easy to fit the ignorable and the nonignorable nonresponse models. We obtain a sample from the joint posterior density in order to make inference about the parameters. Specifically, we need to make inference about  $\mathbf{p}_t$ . To run the Metropolis-Hastings sampler, we need the conditional posterior density of each of the parameters given the others.

First, we consider the conditional posterior probability mass functions of  $\mathbf{y}_{ts}$ ,  $s = 2, 3, 4$ ,  $t = 1, \dots, T$  given  $\mathbf{y}_{1(s)}$ ,  $\mathbf{p}_t$ ,  $\boldsymbol{\pi}_{tjk}$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ . From (14) it is clear that under the conditional posterior density the  $\mathbf{y}_{ts}$ ,  $t = 1, \dots, T$ ,  $s = 2, 3, 4$ , are independent multinomial random vectors. Specifically, letting  $\mathbf{p} = (p_{tjk}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$  and  $\boldsymbol{\pi} = (\pi_{tjk}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ ,

$$\mathbf{y}_{t2j} \mid \{\mathbf{y}_{t1}, \mathbf{p}, \boldsymbol{\pi}\} \stackrel{\text{ind}}{\sim} \text{Multinomial}(u_{tj}, \mathbf{q}_{tj}^{(2)}) \quad j = 1, \dots, r,$$

$$\mathbf{y}_{t3k} \mid \{\mathbf{y}_{t1}, \mathbf{p}, \boldsymbol{\pi}\} \stackrel{\text{ind}}{\sim} \text{Multinomial}(v_{tk}, \mathbf{q}_{tk}^{(3)}) \quad k = 1, \dots, c,$$

$$\mathbf{y}_{t4} \mid \{\mathbf{y}_{t1}, \mathbf{p}, \boldsymbol{\pi}\} \sim \text{Multinomial}(w_t, \mathbf{q}_t^{(4)}), \quad (15)$$

where  $q_{ijk}^{(2)} = \pi_{t2jk} p_{ijk} / \sum_{k=1}^c \pi_{t2jk} p_{ijk}$ ,  $k = 1, \dots, c$ ,  $q_{ijk}^{(3)} = \pi_{t3jk} p_{ijk} / \sum_{j=1}^r \pi_{t3jk} p_{ijk}$ ,  $j = 1, \dots, r$  and  $q_{ijk}^{(4)} = \pi_{t4jk} p_{ijk} / \sum_{j=1}^r \sum_{k=1}^c \pi_{t4jk} p_{ijk}$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ ,  $t = 1, \dots, T$

The conditional posterior density of  $\pi_{ijk}$  is given by

$$\pi_{ijk} \mid \{\boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{y}_t\} \sim \text{Dirichlet}(\nu_{i1jk} + \mu_{i1}\tau_{1i}, \nu_{i2jk} + \mu_{i2}\tau_{1i}, \nu_{i3jk} + \mu_{i3}\tau_{1i}, \nu_{i4jk} + \mu_{i4}\tau_{1i}) \quad (16)$$

with independence over  $t = 1, \dots, T$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ .

The conditional posterior density for  $\boldsymbol{p}_t$ ,  $t = 1, \dots, T$  is more difficult. We note that

$$\pi(\boldsymbol{p}_0 \mid \text{else}, \boldsymbol{y}_1) \propto \frac{\prod_{j=1}^r \prod_{k=1}^c p_{0jk}^{\nu_{0j} \tau_{2j} - 1}}{D(\boldsymbol{p}_0 \tau_2)} \quad (17)$$

and

$$\pi(\boldsymbol{p}_t \mid \text{else}, \boldsymbol{y}_1) \propto \left\{ \prod_{j=1}^r \prod_{k=1}^c p_{tjk}^{\nu_{tjk} + p_{t-1jk} \tau_{2j} - 1} \right\} \frac{\prod_{j=1}^r \prod_{k=1}^c p_{t+1jk}^{\nu_{t+1jk} - 1}}{D(\boldsymbol{p}_t \tau_2)}, t = 1, \dots, T, \quad (18)$$

where “else” refers to all of the parameters in  $(\boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{y}_{(1)})$  excluding  $\boldsymbol{p}_0$  in (17) or  $\boldsymbol{p}_t$  in (18). We show how to draw samples from (17) and (18) in Appendix A.

Next, we consider the hyper-parameters. Letting  $\delta_s = \prod_{t=1}^T \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}$ , and  $\boldsymbol{\pi} = (\pi_{tjk}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ , the joint conditional posterior density of  $\boldsymbol{\mu}_1, \boldsymbol{\tau}_1$  is

$$p(\boldsymbol{\mu}_1, \boldsymbol{\tau}_1 \mid \boldsymbol{\pi}) \propto \frac{\prod_{s=1}^4 \delta_s^{\mu_{1s} \tau_1}}{\{D(\boldsymbol{\mu}_1 \tau_1)\}^{rcT}} \tau_1^{1/c_0^2 - 1} e^{-\tau_1 / \mu_0 c_0^2},$$

where

$$\sum_{s=1}^4 \mu_{1s} = 1, \mu_{1s} \geq 0, s = 1, 2, 3, 4, \tau_1 > 0.$$

We do not need to get a sample directly from  $p(\boldsymbol{\mu}_1 \mid \boldsymbol{\tau}_1, \boldsymbol{\pi})$ . But, letting  $\boldsymbol{\mu}_{1(s)}$  denote the vector of all components of  $\boldsymbol{\mu}_1$  except  $\mu_{1s}$ , we have

$$p(\boldsymbol{\mu}_{1s} \mid \boldsymbol{\mu}_{1(s)}, \boldsymbol{\tau}_1, \boldsymbol{\pi}) \propto \frac{\delta_s^{\mu_{1s} \tau_1}}{\{\Gamma(\mu_{1s} \tau_1)\}^{rcT}} \frac{\delta_4^{(1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1}}{\{\Gamma((1-\mu_{11}-\mu_{12}-\mu_{13})\tau_1)\}^{rcT}},$$

$$0 \leq \mu_{1s} \leq 1 - \sum_{s'=1, s' \neq s}^3 \mu_{1s'}, s = 1, 2, 3. \quad (19)$$

We use a grid method to draw a sample from  $p(\boldsymbol{\mu}_{1s} \mid \boldsymbol{\mu}_{1(s)}, \boldsymbol{\tau}_1, \boldsymbol{\pi})$ . We started by using 50 grids (*i.e.*, we have divided the range of  $\mu_{1s}$ ,  $(0, 1 - \sum_{s'=1, s' \neq s}^3 \mu_{1s'})$ , into 50 intervals of equal widths) to form an approximate probability mass function of  $\mu_{1s}$ ,  $s = 1, 2, 3$ . We first draw a

random variable from this probability mass function to indicate which of the 50 intervals is selected. Then, for  $\mu_{1s}$  we draw a uniform random variable in this interval. This procedure is efficient because  $\mu_{1s}$  is bounded, the intervals are very narrow, and it is very “cheap” to construct the discrete probability mass function for each  $\mu_{1s}$ ,  $s = 1, 2, 3$ . Finally,  $\mu_{14}$  is obtained from its conditional posterior density by taking  $\mu_{14} = 1 - \sum_{s=1}^3 \mu_{1s}$ .

The conditional posterior density of  $\tau_1$  is

$$p(\tau_1 \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \left[ \prod_{s=1}^4 \frac{\delta_s^{\mu_{1s} \tau_1}}{\{\Gamma(\mu_{1s} \tau_1)\}^{rcT}} \right] \tau_1^{1/c_0^2 - 1} e^{-\tau_1 / \mu_0 c_0^2}, \tau_1 > 0. \quad (20)$$

To draw a random deviate from (20), we proceed in the same manner as for (19), except that we transform  $\tau_1$  from the positive half of the real line to  $(0, 1)$ . (It is more convenient to perform a grid approximation to a density in a bounded interval.) Thus, letting  $\tau_1 = \phi / (1 - \phi)$  in (20), we have

$$p(\phi \mid \boldsymbol{\mu}_1, \boldsymbol{\pi}) \propto \frac{1}{(1 - \phi)^2} \left[ \prod_{s=1}^4 \frac{\delta_s^{\mu_{1s} \tau_1}}{\{\Gamma(\mu_{1s} \tau_1)\}^{rcT}} \right] \tau_1^{1/c_0^2 - 1} e^{-\tau_1 / \mu_0 c_0^2} \Big|_{\tau_1 = \frac{\phi}{1-\phi}}, 0 < \phi < 1.$$

Again, we started by using 50 intervals of equal width to draw  $\phi$ , and the random deviate for  $\tau_1$  is  $\phi / (1 - \phi)$ .

Letting  $\boldsymbol{p} = (p_{tjk}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ , the conditional posterior density of  $\tau_2$  is

$$\pi(\tau_2 \mid \boldsymbol{p}) \propto \frac{1}{(1 + \tau_2)^2} \prod_{t=1}^T \left\{ \frac{\prod_{j=1}^r \prod_{k=1}^c p_{tjk}^{p_{t-1jk} \tau_2 - 1}}{D(\boldsymbol{p}_{t-1} \tau_2)} \right\}, \tau_2 > 0. \quad (21)$$

A sample is obtained in a manner similar to  $\tau_1$  in (20).

We have extensive experience in using the grid approximation. However, one has to be careful in using the grid approximation for parameters close to 0 or 1 in the interval  $[0, 1]$ . One would need to use a grid approximation in an interval near the boundary; this can be obtained by trial and error in looking at the output of the sampler as it progresses. If a parameter in  $[0, 1]$  is likely to be away from 0 or 1, then the grid method works fine; this is the case for the  $\mu_{1s}$ 's. However, for a parameter like  $\tau_1$  (can be very large), when transformed to  $\phi$  in the interval  $[0, 1]$ ,  $\phi$  can be very large (near to 1). If the transformed value is like 0.999, one needs to adjust the grid search to be in an interval containing 0.999. This has to be done by trial and error; one needs to look at the output of  $\phi$  as the sampler progresses, and adjust the interval accordingly. For example, if 100 grid points are equally spaced in  $[0, 1]$  such as 0.01, 0.02, 0.03, ..., 0.99, and the parameter is likely to be around 0.999, although we draw uniformly in the selected grid interval, these grid points are not going to be very efficient.

The Metropolis-Hastings sampler is executed by drawing a random deviate from each of (15), (16), (17), (18), (19), (20) and (21) iterating the entire procedure until convergence. This is an example of the griddy Metropolis-Hastings sampler (Ritter and Tanner 1992). We obtain a sample from the posterior densities corresponding to the ignorable and nonignorable nonresponse models in a similar manner. For all models, we use a sample of  $M = 1,000$  from the posterior densities to do estimation and prediction. We monitored the algorithm for convergence by looking at the trace plots of each parameter versus iteration order and we studied the autocorrelation coefficient. We used a griddy Gibbs sampler to fit the ignorable and nonignorable nonresponse models. We used a "burn in" of 1,000 iterates and we took every tenth thereafter. This procedure works well.

However, for the time-dependent model, we used a griddy approximation to the conditional posterior of  $p_0$ , but Metropolis steps for  $p_t, t = 1, \dots, T$ . The Metropolis steps did not work well because the jumping probabilities are 0.67, 0.65 and 0.73 for the three conditional posterior densities of  $p_1, p_2$ , and  $p_3$ , but they are recommended to be between 0.25 and 0.50 (Gelman, Roberts and Gilks 1996); tuning did not help. So we used grid approximations to these three conditional posterior densities as well. The grid approximations are very accurate. In all grid approximations, we started with 50 grids, and we increased the number of grids until our estimates of all  $p_T, p_{T+1}, \mu_2, \tau_2$  do not change. We found that 200 grids were adequate in all cases (*i.e.*, for  $\mu_1, \mu_2, \tau_1, \tau_2$ ). Also, we found that although the Metropolis-Hastings sampler did not work as well as we wanted, the estimates of the cell proportions are virtually the same from both samplers. The Metropolis-Hastings sampler was run for 25,000 iterations with a "burn in" of 5,000 and thinning by choosing every twentieth.

Finally, we stored the sample from the joint posterior density for further analysis. Specifically, for the ignorable and the nonignorable nonresponse models, we need the sample of size  $M$  from  $\{(\mu_2^{(h)}, \tau_2^{(h)}, p_{T-1}^{(h)}, p^{(h)}), h = 1, \dots, M\}$ , and for the time-dependent model we need the sample of size  $M$  from  $(\tau_2^{(h)}, p_{T-1}^{(h)}, p^{(h)}), h = 1, \dots, M$ .

### 3.4 Specification of $\mu_0$ and $c_0^2$

Finally, we describe how to specify  $\mu_0$  and  $c_0$  in (12). This is important because it permits us to "center" the nonignorable nonresponse model on the ignorable nonresponse model (*i.e.*, an expansion model). This procedure is in the spirit of Nandram, *et al.* (2005).

We have drawn a sample of  $\pi_t^{(h)}, t = 1, \dots, T, h = 1, \dots, M$ ,  $M = 1,000$  iterates from the ignorable nonresponse model, and computed  $\pi^{(h)} = \sum_{t=1}^T \pi_t^{(h)} / T, h = 1, \dots, M$ . Then, using the griddy Gibbs sampler, we fit the model

$$\pi^{(h)} \sim \text{Dirichlet}(\mu; \tau_1),$$

$$\mu_1 \sim \text{Dirichlet}(\mathbf{1}), \quad p(\tau_1) = 1/(1 + \tau_1)^2, \quad \tau_1 > 0,$$

with *a priori*  $\mu_1$  and  $\tau_1$  independent, to obtain a sample  $\tau_1^{(h)}, h = 1, \dots, M$ . We have drawn 1,500 iterates with a "burn in" of 500 to get  $M = 1,000$  iterates.

Finally, taking  $a = M^{-1} \sum_{h=1}^M \tau_1^{(h)}$  and  $b = (M-1)^{-1} \sum_{h=1}^M (\tau_1^{(h)} - a)^2$ , we set

$$c_0 = \sqrt{b/a} \text{ and } \mu_0 = a.$$

For the election data, our procedure gives  $c_0 = 0.031$  and  $\mu_0 = 2.431$ . This specification will hold the nonignorable nonresponse model close to the ignorable nonresponse model, thereby providing a possible centering mechanism.

To study sensitivity to the misspecification of the prior density of  $\tau_1$ , we use two constants,  $\kappa_1$  and  $\kappa_2$ , such that *a priori*

$$\tau_1 \sim \text{Gamma}(1/\kappa_1^2 c_0^2, 1/\kappa_1^2 \kappa_2 \mu_0 c_0^2)$$

with varying values of  $\kappa_1$  and  $\kappa_2$ . It is worth noting that  $E(\tau_1) = \kappa_2 \mu_0$  and  $\text{CV}(\tau_1) = \kappa_1 c_0$ ; thus increasing  $\kappa_2$  means increasing  $\tau_1$  which, in turn, means increasing precision *a priori* but not necessarily *a posteriori*. We will study the sensitivity to the specification of  $\kappa_1$  and  $\kappa_2$  when we describe the data analysis.

### 3.5 Estimation and prediction

We show how to improve estimation (*i.e.*, Rao-Blackwellization) in the October poll, and how to do prediction in the November election.

For the ignorable and nonignorable nonresponse models,

$$\begin{aligned} g(p_T | y_1) &= \int g(p_T | \mu_2, \tau_2) \pi(\mu_2, \tau_2 | y_1) d\mu_2 d\tau_2 \\ &\approx \frac{1}{M} \sum_{h=1}^M g(p_T | \mu_2^{(h)}, \tau_2^{(h)}), \end{aligned} \quad (22)$$

where  $p_T | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2)$ , and for the time-dependent model,

$$\begin{aligned} g(p_T | y_1) &= \int g(p_T | p_{T-1}, \tau_2) \pi(p_{T-1}, \tau_2 | y_1) dp_{T-1} d\tau_2 \\ &\approx \frac{1}{M} \sum_{h=1}^M g(p_T | p_{T-1}^{(h)}, \tau_2^{(h)}), \end{aligned} \quad (23)$$

where  $p_T | p_{T-1}, \tau_2 \sim \text{Dirichlet}(p_{T-1} \tau_2)$ .

We obtain (predict) the cell proportions for November as follows. The ignorable or nonignorable nonresponse model, posterior density of  $p_{T+1}$  is

$$\begin{aligned} g(p_{T+1} | y_1) &= \int g(p_{T+1} | \mu_2, \tau_2) \pi(\mu_2, \tau_2 | y_1) d\mu_2 d\tau_2 \\ &\approx \frac{1}{M} \sum_{h=1}^M g(p_{T+1} | \mu_2^{(h)}, \tau_2^{(h)}), \end{aligned} \quad (24)$$



where  $p_{T+1} \mid \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2, \tau_2)$ . For the time-dependent

$$g(p_{T+1} \mid y_1) = \int g(p_{T+1} \mid p_T, \tau_2) \pi(p_T, \tau_2 \mid y_1) d p_T d \tau_2$$
$$\approx \frac{1}{M} \sum_{h=1}^M g(p_{T+1} \mid p_T^{(h)}, \tau_2^{(h)}), \tag{25}$$

where  $p_{T+1} \mid p_T, \tau_2 \sim \text{Dirichlet}(p_T, \tau_2)$ .

Thus, by (22), (23), (24) and (25), estimation and prediction are straight forward. For example, consider the time-dependent model. For estimation, by (24) for each  $h$ , we draw a random deviate  $p_T \mid p_{T-1}, \tau_2^{(h)} \sim \text{Dirichlet}(p_{T-1}, \tau_2^{(h)})$ , denoted by  $p_T^{(h)}, h=1, \dots, M$ . For prediction, by (25) for each  $h$ , we draw a random deviate  $p_{T+1} \mid p_T^{(h)}, \tau_2^{(h)} \sim \text{Dirichlet}(p_T^{(h)}, \tau_2^{(h)})$ , denoted by  $p_{T+1}^{(h)}, h=1, \dots, M$ . Thus, inference about  $p_T$  and  $p_{T+1}$  is made in the usual manner. The procedure is similar for the ignorable and nonignorable nonresponse models.

4. Data analysis

In this section we compare our models with those of Chen and Stasny (2003) and the actual (November election) outcomes. We have introduced a new parameter to help predict the outcome of the election. We also study extensively sensitivity of inference to choices of  $\kappa_1$  and  $\kappa_2$ . Based on our procedure, we have specified the coefficient of variation,  $c_0 = 0.031$ , and the mean,  $\mu_0 = 2.431$ , of the prior distribution of  $\tau_1$ .

In Table 2 we compare inference about the proportions of October voters allocated to the three candidates by our models and those of Chen and Stasny (2003). In this table the results are based on the prior  $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2)$  (*i.e.*,  $\kappa_1 = \kappa_2 = 1$ ). We also present the actual proportions taken from Chang and Krosnick (2001). The actual proportions are (0.45, 0.50, 0.05) for Fisher, Taft and other. Using our time-dependent nonresponse model these proportions are estimated to be (0.41, 0.50, 0.09). These compare favorably with the actual outcomes. The corresponding estimates are (0.41, 0.51, 0.08) for the ignorable nonresponse model and (0.40, 0.50, 0.09) for the nonignorable nonresponse model. The best result of Chen and Stasny (2003) is obtained from their Model D, and their estimates are (0.42, 0.51, 0.07). We have provided 95% credible intervals for our estimates, but within the approach of Chen and Stasny (2003) it is relatively more difficult to provide similar intervals. Also, in Table 2 we present estimates of the predicted proportions for the November elections. The point predictors are similar to the point estimates except for the predicted proportion going to Taft under the ignorable nonresponse model. However, as

expected the 95% credible intervals for the predicted proportions are much wider. For example, under the time-dependent model 95% credible interval for the proportion voting for Taft in the October poll is (0.41, 0.60) and for prediction it is (0.21, 0.78). Thus, while the point estimates and predictions do indicate the winner, the variability indicates no difference between Taft and Fisher. We will look at this further.

**Table 2**  
**Comparison of the proportion of likely voters for the October 1998 poll and prediction for November 1998 election for different models with actual outcome**

Status	Fisher	Taft	Other
Sample Estimate	0.41	0.51	0.08
Approximate 95% CI	(0.35, 0.47)	(0.45, 0.57)	(0.05, 0.11)
Actual Outcome	0.45	0.50	0.05
a. Estimation			
Chen/Stasny models A,B,C	0.41	0.51	0.08
Chen/Stasny model D	0.42	0.51	0.07
Chen/Stasny model E	0.41	0.51	0.08
Ignorable model	0.41	0.51	0.08
95% CI	(0.35, 0.46)	(0.46, 0.57)	(0.05, 0.12)
Nonignorable model	0.41	0.50	0.09
95% CI	(0.32, 0.51)	(0.40, 0.60)	(0.05, 0.17)
Time-dependent model	0.41	0.50	0.09
95% CI	(0.32, 0.52)	(0.41, 0.60)	(0.05, 0.16)
b. Prediction			
Ignorable model	0.41	0.54	0.05
95% CI	(0.15, 0.70)	(0.25, 0.81)	(0.00, 0.22)
Nonignorable model	0.42	0.52	0.06
95% CI	(0.15, 0.70)	(0.22, 0.79)	(0.00, 0.28)
Time-dependent model	0.41	0.50	0.09
95% CI	(0.15, 0.71)	(0.21, 0.78)	(0.00, 0.31)

NOTE:  $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2)$ , where  $c_0 = 0.031$  and  $\mu_0 = 2.431$ .

Although our estimates from the time-dependent model are close to the actual estimates, the 95% credible intervals for  $p_{311}$  and  $p_{312}$  overlap, thereby making it difficult to predict Taft is the winner. Although the 95% credible intervals for our other models are shorter, the point estimates are not so good and they still overlap. One weakness in our analysis in Table 2 is that we have ignored the correlation between the two estimates (*i.e.*, we should really study the difference  $p_{312} - p_{311}$ , the margin of winning).

In Table 3 we present estimates of  $\Lambda_e = p_{312} - p_{311}$  and  $\Lambda_p = p_{412} - p_{411}$  at  $\kappa_1 = \kappa_2 = 1$  for the three models. We have also included the numerical standard error (NSE) which is a measure of how well the numerical results can be reproduced; we have used the batch-means method to compute it. Small NSEs mean that if we repeat the entire computation the same way (*i.e.*, using another 1,000 iterates), we should see very little difference between the two sets of answers. In Table 3 the NSEs are small. The point estimators and predictors are all positive showing that Taft is the winner in both the October poll and the November election. However, the variability dwarfs this

result somewhat because the PSD are large, as expected even more so for prediction. This causes the 95% credible intervals for both parameters to contain 0. Thus, again when variability is considered, there is no difference between Taft and Fisher.

Table 3

Comparison of the three models for estimation and prediction using the posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals for  $\Lambda_e$  ( $\Lambda_p$ ) and  $\Delta_e$  ( $\Delta_p$ )

Model	PM	PSD	NSE	Interval
$\Lambda_e$	Ignorable	0.105	0.055	0.002 (-0.002, 0.209)
	Nonignorable	0.097	0.099	0.006 (-0.100, 0.280)
	Time-dependent	0.093	0.101	0.007 (-0.098, 0.276)
$\Lambda_p$	Ignorable	0.071	0.154	0.004 (-0.240, 0.362)
	Nonignorable	0.058	0.150	0.005 (-0.252, 0.369)
	Time-dependent	0.050	0.134	0.005 (-0.244, 0.314)
$\Delta_e$	Ignorable	0.688	0.175	0.008 (0.295, 0.958)
	Nonignorable	0.663	0.200	0.012 (0.222, 0.959)
	Time-dependent	0.632	0.148	0.014 (0.336, 0.901)
$\Delta_p$	Ignorable	0.688	0.175	0.008 (0.295, 0.960)
	Nonignorable	0.663	0.193	0.009 (0.253, 0.972)
	Time-dependent	0.648	0.155	0.011 (0.341, 0.923)

NOTE: See note to Table 2;  $\Lambda_e = p_{312} - p_{311}$  (estimation, difference between Taft and Fisher for the October poll);  $\Lambda_p = p_{412} - p_{411}$  (prediction, difference between Taft and Fisher for the November election);  $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \alpha)$ ; and  $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \alpha)$ ; see (26).

We seek an alternative parameter looking to help us predict the winner more convincingly. We pose the following question: "What is the probability that the proportion of Taft's voters in the October poll and the November election is larger than that of Fisher's voters?"

Thus, we consider the parameter  $\Delta_e = \Pr(p_{312} > p_{311} | p_{311} + p_{312} + p_{313}, \alpha)$  where  $\alpha_{jk} = \mu_{jk} \tau_2$ ,  $j=1, \dots, r$ ,  $k=1, \dots, c$ , for the ignorable and nonignorable nonresponse models, and  $\alpha_{jk} = p_{2jk} \tau_2$ ,  $j=1, \dots, r$ ,  $k=1, \dots, c$ , for the time-dependent model. In either case, letting  $q_1 = p_{311}/p_{31\cdot}$ ,  $q_2 = p_{312}/p_{31\cdot}$ , and  $q_3 = p_{313}/p_{31\cdot}$  with  $p_{31\cdot} = \sum_{k=1}^3 p_{31k}$  and  $\sum_{k=1}^3 q_k = 1$ , it is easy to show that  $(q_1, q_2, q_3) \sim \text{Dirichlet}(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)$ , where  $\tilde{\alpha}_1 = \alpha_{11}$ ,  $\tilde{\alpha}_2 = \alpha_{12}$  and  $\tilde{\alpha}_3 = \alpha_{13} + \sum_{k=1}^c \alpha_{2k}$ . Therefore, we have

$$\Delta_e = \Pr(q_2 > q_1 | \alpha)$$

$$= \int_0^{1/2} \left\{ \int_{q_1}^{1-q_1} \frac{q_1^{\tilde{\alpha}_1-1} q_2^{\tilde{\alpha}_2-1} (1-q_1-q_2)^{\tilde{\alpha}_3-1}}{D(\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3)} dq_2 \right\} dq_1.$$

Then, it is easy to show that

$$\Delta_e = 1 - F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(1/2) \int_0^{1/2} F_{\tilde{\alpha}_2, \tilde{\alpha}_3}\{q_1/(1-q_1)\} \left\{ \frac{q_1^{\tilde{\alpha}_1-1} (1-q_1)^{\tilde{\alpha}_2 + \tilde{\alpha}_3 - 1}}{F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(1/2) B(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)} \right\} dq_1, \quad (26)$$

where

$$F_{\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_1-1} (1-x)^{\tilde{\alpha}_2 + \tilde{\alpha}_3 - 1}}{B(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)} dx$$

and

$$F_{\tilde{\alpha}_2, \tilde{\alpha}_3}(a) = \int_0^a \frac{x^{\tilde{\alpha}_2-1} (1-x)^{\tilde{\alpha}_3-1}}{B(\tilde{\alpha}_2, \tilde{\alpha}_3)} dx.$$

We note that  $\Delta_e$  is the probability that Taft received a higher proportion of the votes in the October poll, and  $\Delta_p$  is the probability that Taft received a higher proportion of the votes in the November election. These parameters can be very useful for estimation ( $e$ ) and prediction ( $p$ ). Parameters like  $\Delta_e$  or  $\Delta_p$  are difficult to analyze in the non-Bayesian approach such as that of Chen and Stasny (2003); indeed this is a great strength of the Bayesian paradigm.

It is easy to compute (26) using Monte Carlo integration. For each  $\tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3$ ,  $\Delta_e q_1 \sim \text{Beta}(\tilde{\alpha}_1, \tilde{\alpha}_2 + \tilde{\alpha}_3)$  truncated to the  $(0, 1/2)$  is used as an importance function. Thus, for each  $\tilde{\alpha}^{(h)}$ ,  $h=1, \dots, M$ ,  $M=1,000$  from the Metropolis-Hastings sampler (or Gibbs sampler), we can compute  $\Delta_e^{(h)}$ . *A posteriori* inference about  $\Delta_e$  is obtained in the standard empirical manner. For prediction, we have also considered  $\Delta_p = \Pr(p_{412} > p_{411} | p_{411} + p_{412} + p_{413}, \alpha)$ , where  $\alpha_{jk} = \mu_{jk} \tau_2$ ,  $j=1, \dots, r$ ,  $k=1, \dots, c$ , for the ignorable and nonignorable nonresponse models, and  $\alpha_{jk} = p_{3jk} \tau_2$ ,  $j=1, \dots, r$ ,  $k=1, \dots, c$ , for the time-dependent model. Note that  $\Delta_e$  and  $\Delta_p$  are the same for the ignorable and nonignorable nonresponse models.

In Table 3 we also present estimates of  $\Delta_e$  and  $\Lambda_p$  for the three models. First, note again that the NSEs are all small. The estimates of these parameters are similar for the three models, and larger than 0.60, but the 95% credible intervals contain 0.5. Thus, again the posterior means indicate that Taft is the winner, but variation is nullifying the effect of Taft being the winner. We note again that the time-dependent model provides sharper inference, not enough though. The parameters  $\Delta_e$  and  $\Delta_p$  are more sensible because they restrict inference to a smaller region by conditioning on  $p_{311} + p_{312} + p_{313}$  and  $p_{411} + p_{412} + p_{413}$ , and from a probabilistic view these parameters are more appropriate.

Finally, we study sensitivity to inference about  $\Lambda_p$  and  $\Delta_p$  for the nonignorable nonresponse model and the time-dependent model. We do not present results for  $\Lambda_e$  and  $\Delta_e$  because they are similar to  $\Lambda_p$  and  $\Delta_p$ . Also, we have dropped the ignorable model as well, and we do not present 95% credible intervals because the posterior densities are roughly symmetric. Our results are presented in Table 4 by model,  $\kappa_1$  and  $\kappa_2$ . The posterior means of  $\Lambda_p$  and  $\Delta_p$  are respectively very similar for different values of  $\kappa_1$  and  $\kappa_2$ . Note that *a priori*

**Table 4**  
**Sensitivity of the posterior means (PM) and the posterior standard deviations (PSD) of  $\Lambda_p$  and  $\Delta_p$  with respect to changes in  $\kappa_1$  and  $\kappa_2$  by model**

Model	$\kappa_1$	$\kappa_2$							
		1		5		25		50	
		PM	PSD	PM	PSD	PM	PSD	PM	PSD
a. $\Lambda_p$									
Nonignorable	1	0.058	0.150	0.046	0.153	0.060	0.148	0.054	0.147
	2	0.051	0.153	0.046	0.146	0.062	0.151	0.054	0.145
	3	0.058	0.152	0.059	0.145	0.053	0.149	0.055	0.149
	4	0.055	0.151	0.057	0.148	0.049	0.148	0.043	0.154
Time-dependent	1	0.050	0.134	0.044	0.144	0.048	0.136	0.050	0.130
	2	0.049	0.136	0.052	0.140	0.056	0.129	0.047	0.137
	3	0.039	0.139	0.049	0.137	0.045	0.139	0.052	0.133
	4	0.037	0.138	0.042	0.138	0.041	0.141	0.051	0.129
b. $\Delta_p$									
Nonignorable	1	0.663	0.200	0.650	0.194	0.666	0.186	0.670	0.182
	2	0.663	0.197	0.661	0.188	0.667	0.185	0.659	0.181
	3	0.663	0.199	0.647	0.196	0.666	0.184	0.669	0.180
	4	0.641	0.202	0.668	0.191	0.643	0.197	0.650	0.195
Time-dependent	1	0.648	0.155	0.642	0.123	0.657	0.099	0.661	0.095
	2	0.660	0.151	0.652	0.127	0.659	0.102	0.657	0.099
	3	0.622	0.153	0.636	0.137	0.649	0.120	0.648	0.115
	4	0.610	0.162	0.636	0.152	0.646	0.132	0.644	0.127

NOTE: We have taken  $\tau_1 \sim \text{Gamma}(1/\kappa_1^2 c_0^2, 1/\kappa_2 \mu_0 \kappa_1^2 c_0^2)$  and we studied sensitivity with respect to  $\kappa_1$  and  $\kappa_2$ . See note to Table 3.

$$\tau_1 \sim \text{Gamma}\left(\frac{1}{\kappa_1^2 c_0^2}, \frac{1}{\kappa_2 \mu_0 \kappa_1^2 c_0^2}\right),$$

$E(\tau_1) = \kappa_2 \mu_0$  and  $SD(\tau_1) = \kappa_1 \kappa_2 c_0 \mu_0$ ; so clearly, *a priori*  $E(\tau_1)$  increases with  $\kappa_2$  and  $SD(\tau_1)$  increases with either  $\kappa_1$  or  $\kappa_2$ , but not necessarily *a posteriori*. These changes do not have a lot of effect on inference *a posteriori*. For almost all combinations of  $\kappa_1$  and  $\kappa_2$ , under the time-dependent model posterior standard deviations of  $\Lambda_p$  are smaller (but not substantially) than under the nonignorable nonresponse model. Under the time-dependent model posterior standard deviations of  $\Delta_p$  are substantially smaller than under the nonignorable nonresponse model for all combinations of  $\kappa_1$  and  $\kappa_2$ .

**Concluding Remarks**

The main contribution in this paper is the construction and analysis of a time-dependent nonignorable nonresponse model and its application to the Ohio polling data. We have done two additional things as well. First, we have compared (the time-dependent model with an extended version (to include time) of the ignorable and nonignorable nonresponse models of Nandram, *et al.* (2005). Second, we have constructed a new parameter to help predict the winner; however, this parameter did not make an enormous difference partly because there are only three time points in the time-dependent model.

Our time-dependent model provides posterior inferences that are closer to the truth than the ignorable and nonignorable nonresponse models as well as those of Chen and Stasny (2003). It is natural for voters' preference to change as new information, detrimental or supportive, is revealed into the public place. Thus, our time-dependent model, which takes care of changes over time and provides improved precision, is to be preferred. The uncertainty in the prediction can be reduced in two ways. First, with an increased number of polls there will be increased precision in the parameters, which in turn, can lead to improved prediction. Second, with more prior information (*e.g.*, exit polling) about the November election, one can also improve the prediction.

Our 95% credible intervals can be shortened by using prior information on the proportion of voters going to Taft or Fisher. A referee suggested, "The major-party voting proportions are between 35% and 65% in general elections, and in specific states an objective political scientist could generally provide an even tighter prior." However, this is a complex problem because with truncated prior distributions on the  $p$ s, there is a normalization constant which is a function of  $\tau_2$ . Thus, when  $\tau_2$  is drawn from its conditional posterior density, we need to perform a Monte Carlo integration to compute the normalization constant at each iterate. While this will be a useful contribution, we prefer not to pursue this problem here.



The number of days to an election has an important impact on poll accuracy and that this effect can vary substantially across different campaign contexts (e.g., DeSart and Holbrook 2003). Thus, it is really difficult to predict the outcome of an election weeks before it actually occurs, unless there exists an absolute margin. Someone who wishes to predict the outcome of an election must take into consideration additional information near the actual election. Our prediction assumes that there is no catastrophic change near the election; such an abrupt change in public opinion can occur. For example, in 1988 Dukakis lost the election against George Bush for various reasons: he spent the last week in Massachusetts, his cold personality, and Bush's attack on his liberal position. Also, an effective campaign can mobilize undecided voters near the election (e.g., Truman and Dewey in 1948). One way to capture a possible catastrophe is to use mixture distributions or other heavy-tailed distributions (as researchers use Levy distributions in mathematical finance).

### Acknowledgement

This work was done while Balgobin Nandram was on sabbatical leave at the National Center for Health Statistics, Hyattsville, Maryland, 2003-2004.

### Appendix A

#### Time-dependent model: Conditional posterior densities of $p_t, t = 0, \dots, T$

We show how to draw a sample from the conditional posterior density of  $p_0$  in (17) using a grid method, and how to draw a sample from the conditional posterior densities of  $p_t, t = 1, \dots, T$  in (18) using Metropolis steps, each with an independence chain.

First, we show how to draw a sample from the conditional posterior density of  $p_0$  in (17) using a grid method. Letting  $(q_{01}, \dots, q_{0L}) = (p_{01}, \dots, p_{0rc})$  and  $(q_{11}, \dots, q_{1L}) = (p_{11}, \dots, p_{1rc})$  where  $L = rc$ , with  $\sum_{\ell=1}^{L-1} q_{0\ell} \leq 1$ , we have

$$\pi(q_{01}, \dots, q_{0L-1} \mid \text{else}, y_1) \propto \frac{q_{1L} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{\tau_2-1}}{\Gamma((1 - \sum_{\ell=1}^{L-1} q_{0\ell}) \tau_2)} \prod_{\ell=1}^{L-1} \frac{q_{1\ell}^{q_{0\ell} \tau_2-1}}{\Gamma(q_{0\ell} \tau_2)}, 0 \leq q_{0\ell} \leq 1, \ell = 1, \dots, L-1,$$

and it is easy to show that

$$\pi(q_{0\ell} \mid \text{else}, y_1) \propto \frac{q_{1L} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{\tau_2-1}}{\Gamma((1 - \sum_{\ell=1}^{L-1} q_{0\ell}) \tau_2)} \frac{q_{1\ell}^{q_{0\ell} \tau_2-1}}{\Gamma(q_{0\ell} \tau_2)},$$

$$0 \leq q_{0\ell} \leq 1 - \sum_{\ell=1, \ell \neq \ell}^{L-1} q_{0\ell}, \ell = 1, \dots, L-1.$$

For each  $\ell$  we divide the range  $0 \leq q_{0\ell} \leq 1 - \sum_{\ell=1, \ell \neq \ell}^{L-1} q_{0\ell}$  into a number of subintervals. To obtain a random deviate  $q_{0\ell}$  from its conditional posterior density, we select an interval proportional to its area, and draw a uniform random deviate from this interval.

Second, we show how to draw a sample from the conditional posterior densities of  $p_t, t = 1, \dots, T$  in (18) using Metropolis steps, each with an independence chain. Consider  $p_t \mid p_{t-1}, \tau_2, y, t = 1, \dots, T$ . We use the candidate generating density

$$p_t \mid p_{t-1}, \tau_2, y \sim \text{Dirichlet}(a_t),$$

where

$$a_{ijk} = y_{t-ijk} + \tau_2 p_{t-1jk}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c.$$

Then, the acceptance probability is  $A_{s,s+1} = \min(1, \psi_{s+1}/\psi_s)$  where

$$\psi_s = \prod_{j=1}^r \prod_{k=1}^c p_{t+1jk}^{p_{jk}^{(s)} \tau_2-1} / D(p^{(s)} \tau_2).$$

### References

- Chang, L.C., and Krosnick, J.A. (2001). Improving election forecasting. *Technical Report*, Department of Psychology, The Ohio State University.
- Chen, Q.L., and Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.
- Chen, T., and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.
- DeSart, J., and Holbrook, T.M. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.
- Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.
- Gelman, A., Roberts, G.O. and Gilks, W.R. (1996). Efficient Metropolis jumping rules. In *Bayesian Statistics*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith), Oxford, U.K.: Oxford University Press, 599-607.

- Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. New York: Chapman & Hall.
- Kalton, G., and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, 12, 1-16.
- Lavrakas, P.J. (1993). *Telephone Survey Methods: Sampling, Selection, and Supervision*. Newbury Park, CA: Sage Publications.
- Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. 2<sup>nd</sup> edition, New York: John Wiley & Sons Inc.
- Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.
- Nandram, B., and Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.
- Nandram, B., and Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.
- Nandram, B., Cox, L.H. and Choi, J.W. (2005). Bayesian analysis of nonignorable missing categorical data: An application to bone mineral density and family income. *Survey Methodology*, 31, 213-225.
- Ritter, C., and Tanner, M.A. (1992). The Gibbs stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.
- Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.





# A noninformative Bayesian approach to finite population sampling using auxiliary variables

Radu Lazar, Glen Meeden and David Nelson<sup>1</sup>

## Abstract

In finite population sampling prior information is often available in the form of partial knowledge about an auxiliary variable, for example its mean may be known. In such cases, the ratio estimator and the regression estimator are often used for estimating the population mean of the characteristic of interest. The Polya posterior has been developed as a noninformative Bayesian approach to survey sampling. It is appropriate when little or no prior information about the population is available. Here we show that it can be extended to incorporate types of partial prior information about auxiliary variables. We will see that it typically yields procedures with good frequentist properties even in some problems where standard frequentist methods are difficult to apply.

Key Words: Sample survey; Noninformative Bayes; Auxiliary variable; Linear constraints and Polya posterior.

## 1. Introduction

Finite population sampling is one area of statistics where prior information is used routinely when making inferences. In most cases this prior information is not incorporated into the problem in a Bayesian manner. One reason for this is that the prior information usually does not lead, in a obvious way, to a sensible prior distribution. In the predictive approach (see Valliant, Dorfman and Royall 2000) a model is assumed and its unknown parameters are estimated after the sample has been observed. In the usual frequentist theory the prior information is encapsulated in the probability sampling plan or sample design. Basu showed that after the sample has been observed, the sampling design plays no role in the posterior distribution for a Bayesian. (For this fact and more of Basu's thoughts on finite population sampling see Ghosh (1988).) Although this radical conclusion has not been accepted by all Bayesians it is true that the usual frequentist theory and the Bayesian approach to survey sampling do not have a lot in common.

Traditional theory in survey sampling has emphasized estimation of the population mean. When the population mean of an auxiliary variable is known *a priori* the ratio estimator or the regression estimator is often employed. If one wishes to estimate something other than the mean, say a population quantile or the population distribution function, or if one has prior information about the auxiliary variable other than its mean then new methods need to be developed. Recent work along this line can be found in Chen and Qin (1993), Chen and Sitter (1999), Mak and Kuk (1993), Kuk and Mak (1989), Rao, Kovar and Mantel (1990) and Chambers and Dunstan (1986).

One advantage of a Bayesian approach is that a sensible posterior distribution for the population will incorporate the prior information into the estimation of several population parameters. Even if the posterior does not have a closed expression for a particular estimator for any given sample we can find its value approximately. This is done by sampling from the posterior distribution to simulate complete copies of the population and employing Monte Carlo estimation methods. If the posterior does not have a convenient form for sampling one should be able to use Markov Chain Monte Carlo methods to implement the simulation process. For each such simulated copy one computes the value of the parameter of interest. By simulating many such full copies of the population one can find, approximately, the corresponding Bayes point and interval estimates of the given population parameter. The problem then is to find a sensible Bayesian population model which utilizes the type of prior information available for the auxiliary variable.

Often, sensible Bayesian models can be based on the Polya posterior. The Polya posterior is a noninformative Bayesian approach to finite population sampling which uses little or no prior information about the population. A good source for more discussion on this approach is Ghosh and Meeden (1997). It is appropriate when a classical survey sampler would be willing to use simple random sampling as their sampling design. Here we show how it can be extended to cases where prior information about an auxiliary variable is present. For example the mean or median of an auxiliary variable might be known exactly or known to belong to some interval of possible values.

1. Radu Lazar, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: radu@stat.umn.edu; Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: glen@stat.umn.edu; David Nelson, Center for Chronic Disease Outcomes Research, Minneapolis VA Medical Center, Minneapolis, MN 55417. E-mail: Dave.Nelson@med.va.gov.

The basic idea is to constrain or restrict the Polya posterior to put positive probability only on simulated populations which satisfy the constraints specified by the prior information for the auxiliary variables. This appropriately constrained Polya posterior can then be used to make inferences about the population parameters of interest. In Nelson and Meeden (1998) the authors considered several scenarios where a constrained Polya posterior yielded sensible frequentist results. There it was assumed that information about the population quantiles of the auxiliary variable was known *a priori*. Here we assume that we have more general forms of prior information about either the population mean or population quantiles for a set of auxiliary variables. These quantities may either be known exactly or known only to belong to some interval.

In section 2 we review the Polya posterior. In section 3 we introduce the constrained Polya posterior and discuss how to use Markov Chain Monte Carlo methods to find, approximately, the value of an estimate based on this distribution. In section 4 we apply the Polya posterior in a variety of situations and compare it to standard methods. In section 5 we discuss more formally how it relates to some standard frequentist methods. In section 6 we see that the constrained Polya posterior can be used with designs other than simple random sampling. In section 7 we give a few concluding remarks. In the appendix we prove an admissibility result for the constrained Polya posterior which gives a theoretical justification for the methods presented here.

## 2. The Polya posterior

Consider a finite population consisting of  $N$  units. For unit  $i$  let  $y_i$ , a real number, be the unknown value of some characteristic of interest. We assume the unknown state of nature,  $y = (y_1, \dots, y_N)$ , belongs to some known subset of  $N$ -dimensional Euclidean space. Suppose we wish to estimate some function  $\gamma(y)$ , of the unknown state of nature. The next step for a proper Bayesian analysis would be to specify a prior distribution over the parameter space. Then, given a sample generated by the sampling design, one would determine the posterior distribution of the unobserved members of the population conditioned on the values of the observed units in the sample. In most cases the posterior will not depend on the sampling design.

The Polya posterior can be used like a proper posterior distribution although it does not arise from a proper Bayesian model. It would be appropriate when there is little known about the population and the sample is assumed to be representative of the population. An example when it would be appropriate is when the sampling design is simple

random sampling. Next, we briefly describe this distribution and outline its theoretical justification.

Given the data, the Polya posterior is a predictive joint distribution for the unobserved units in the population conditioned on the values in the sample. Given a sample we now show how to generate a set of possible values for the unobserved units from this distribution. Consider two urns where the first urn contains the  $n$  units in the sample along with their observed  $y$  values. The second urn contains the  $N - n$  unsampled units. We begin by choosing one unit at random from each of the two urns. We then assign the observed  $y$  value of the unit selected from the first urn to the unit selected from the second urn and then place them both in the first urn. The urns now contain  $n + 1$  and  $N - n - 1$  balls respectively. This process is repeated until all the units have been moved from the second urn to the first and have been assigned a value. At each step in the process all the units in the first urn have the same probability of being selected. That is, the units which have been assigned a value are treated just like the ones that actually appeared in the sample. Once this is done, we have generated one complete realization of the population under the Polya posterior distribution. This simulated, completed copy contains the  $n$  observed values along with the  $N - n$  simulated values for the unobserved members of the population. Hence, simple Polya sampling yields a predictive distribution for the unobserved given the observed. A good reference for Polya sampling is Feller (1968). The Polya posterior is related to the Bayesian bootstrap of Rubin (1981). See also Lo (1988) and Binder (1982).

This predictive distribution often generates estimators similar to standard frequentist estimators under simple random sampling. Consider, for example, estimation of the population mean. Before continuing we need a bit more notation.

Let  $s$  denote a possible sample of size  $n(s)$ . It is a subset of  $\{1, 2, \dots, N\}$ , the set of labels for the finite population. If  $s = \{i_1, \dots, i_{n(s)}\}$  then  $y_s = \{y_{i_1}, \dots, y_{i_{n(s)}}\}$  is the set of observed values for  $y$ , the characteristic of interest. We let  $z = (s, y_s)$  denote a typical observed sample. Then given  $z = (s, y_s)$  we have

$$\bar{z}_s = \sum_{j=1}^{n(s)} y_{i_j} / n(s)$$

and

$$\text{Var}(z) = \sum_{j=1}^{n(s)} (y_{i_j} - \bar{z}_s)^2 / (n(s) - 1)$$

are the sample mean and sample variance. Let  $\gamma_{mn}(y) = \sum_{i=1}^N y_i / N$  be the population mean. Under the Polya posterior distribution,

$$E(\gamma_{mn}(y) | z) = \bar{z}_s$$



and

$$\text{Var}(\gamma_{mn}(y) | z) = (1 - f) \frac{\text{Var}(z)n(s) - 1}{n(s) \ n(s) + 1}$$

where  $f = n(s)/N$ . Note that, except for the last factor in the posterior variance, these two terms are just the sample mean and its variance under simple random sampling. The design probabilities play no explicit role in these calculations. Nonetheless, for the Polya posterior to be appropriate, in the judgment of the survey sampler, the values for the characteristic of interest for the observed and unobserved units need to be roughly exchangeable. It is in such situations that simple random sampling without replacement is used.

Under the Polya posterior the Bayesian credible interval for the population mean or point and interval estimates of other population quantities cannot always be found explicitly. In such cases it is easy to find these estimates approximately by repeatedly simulating completed copies of the population. For each simulated copy we calculate the population parameter of interest. Experience has shown that 500 to 1,000 simulated values will usually give good results. The mean of these computed values will be our point estimate and the 0.025 and 0.975 quantiles of these computed values will be our interval estimate.

Since under the Polya posterior the only  $y$  values that appear in a simulated completed copy of the population are those that appeared in the sample the Polya posterior is just a way to assign random weights, *i.e.*, probabilities, to the units in the sample. Under the Polya posterior the average weight assigned to each unit in the sample is  $1/n(s)$  so, as we have seen, its estimate of the population mean is just the sample mean. It is this relationship and the Bayes like character of the Polya posterior which allows one to prove the admissibility of the sample mean for estimating the population mean under squared error loss. This suggests that inferential procedures based on the Polya posterior will tend to agree with frequentist procedures and will have good frequentist properties.

As further documentation of this point we note that recently two of the authors (Nelson and Meeden 2006) demonstrated that Bayesian credible intervals based on the Polya posterior for the population median agree asymptotically with the standard Woodruff interval (Woodruff 1952). For another example consider estimating either the mean or the total of a subpopulation or domain when a simple random sample from the entire population is used. Here the number of units in the sample which belong to the domain is a random variable. Hence the mean of the units in the sample which fall into the domain is the ratio of two random variables. This estimate is more complicated than the mean of all the units in the sample. To get an estimate of variance

for this estimator the usual frequentist method conditions on the number of units in the sample that are in the domain. However when estimating the domain total this conditional argument does not work and an unconditional method is used to get an estimate of variance. See for example Cochran (1976). Recently one of the authors (Meeden 2005) showed that inferences based on the Polya posterior agree with the usual frequentist answers. Hence the Polya posterior handles both situations with one simple theory. It is important to remember that conditioning in the frequentist approach can be done under simple random sampling but for more complex designs, conditioning is not generally feasible since the conditional randomization distribution is unknown. As a final example note that the usual frequentist two stage cluster sampling procedures can be been justified from an extension of the Polya posterior (Meeden 1999).

The Polya posterior is similar in spirit to bootstrap methods for finite population sampling. Both methods use a type of exchangeability argument to generate pseudo-versions of the population. The basic idea for the bootstrap is found in Gross (1980). Suppose we have a simple random sample of size  $n(s)$  from the population and suppose  $N/n(s) = m$  is an integer. Given the sample we create a good guess for the population by combining  $m$  replicates of the sample. We then take repeated random samples of size  $n(s)$  from this created population to study the behavior of the estimator of interest. The asymptotic properties of estimators can also be studied (see Booth, Bulter and Hall 1994 for details). This is in contrast to the Polya posterior which for a fixed sample generates complete versions of the population and examines the distribution of the parameter of interest in the population rather than properties of the estimator for the parameter. For the given population quantity of interest the properties of its estimator derive directly from this predictive distribution for the population values.

The Polya posterior is the Bayesian bootstrap of Rubin (1981) applied to finite population sampling. The original Bayesian bootstrap applies to a random sample from an infinite population. Rubin showed that the bootstrap and Bayesian bootstrap are operationally very similar. The same type of analogy holds for the finite population setup. To study the variability of an estimator each repeatedly assigns random weights to the units in the sample. The logic for assigning the weights are different in the two cases as well as their theoretical justifications. The bootstrap has an asymptotic justification under repeated random sampling. The Polya posterior has a decision theoretic justification based on its stepwise Bayes nature (Ghosh and Meeden 1997).

Rather than generating a complete copy of the population it is often more efficient to use a well known approximation



to the Polya posterior. Assume that the sampling fraction  $f$  is small. For  $j = 1, \dots, n(s)$  let  $p_j$  be the proportion of units in a complete simulated copy of the entire population which take on the value  $y_j$ . Then, under the Polya posterior,  $p = (p_1, \dots, p_{n(s)})$  has approximately a Dirichlet distribution with a parameter vector of all ones, *i.e.*, it is uniform on the  $n(s) - 1$  dimensional simplex, where  $\sum_{j=1}^{n(s)} p_j = 1$ . This approach will be very useful when we consider the constrained Polya posterior.

### 3. The constrained Polya posterior

#### 3.1 The basic idea

In many situations, in addition to the variable of interest,  $y$ , the sampler has in hand auxiliary variables,  $x$ , for which prior information is available. For example, the population mean,  $\mu_x$ , of  $x$  could be known. Given a unit in a random sample we observe its pair of values  $(y, x)$ . Following our earlier notation we denote the sample by

$$z = (s, (y, x)_s) = (s, \{(y_i, x_i), \dots, (y_{n(s)}, x_{n(s)})\}).$$

In this situation the regression estimator is often used when estimating the population mean. How should the Polya posterior be adjusted to take into account the fact that the population mean of  $x$  is known? The simple answer is to constrain the predictive distribution to put mass only on populations consistent with the prior information. In practice, we would only generate completed copies of the population consistent with the known prior information. To see how this can be done we consider the approximate form of the Polya posterior described at the end of the previous section.

For  $j = 1, \dots, n(s)$  let  $p_j$  be the proportion of units in a completed copy of the population that have the value  $(y_j, x_j)$ . Rather than using the uniform distribution for  $p = (p_1, \dots, p_{n(s)})$  over the simplex to generate simulated copies of the population we should use the uniform distribution restricted to the subset of the simplex satisfying

$$\sum_{j=1}^{n(s)} p_j x_{ij} = \mu_x. \quad (1)$$

Before describing how we can generate vectors of  $p$  from this constrained Polya posterior we consider how the resulting estimator is related to the regression estimator.

Numerous simulation results (not presented here) show that the constrained Polya posterior behaves very much like the regression estimator under simple random sampling. The following simple argument shows why these two point estimates should often agree even though the Polya posterior makes no assumptions about the relationship between  $y$  and  $x$ .

Suppose in the population  $y_i = a + bx_i + \varepsilon_i$  where  $\varepsilon_i$  is a random error with expectation zero. Let  $\bar{X}$  be the known population mean of  $x$ . Then given a sample and  $p_i$ 's satisfying  $\sum_{i \in s} p_i x_i = \bar{X}$  we have

$$\begin{aligned} E\left(\sum_{i \in s} p_i y_i\right) &= aE\left(\sum_{i \in s} p_i\right) + bE\left(\sum_{i \in s} p_i x_i\right) + \sum_{i \in s} E(p_i \varepsilon_i) \\ &\doteq a + b\bar{X} \\ &\doteq \bar{y}_s - \hat{b}\bar{x}_s + \hat{b}\bar{X} \\ &= \bar{y}_s + \hat{b}(\bar{X} - \bar{x}_s) \end{aligned}$$

where  $\hat{b}$  is the least squares estimate of  $b$ . Here the sample values are fixed and the  $p_i$ 's and  $\varepsilon_i$ 's are random and the expectation of the  $p_i$ 's is with respect to the constrained Polya posterior. The first approximation follows since under simple random sampling we expect to see balanced samples on the average and the  $p_i$ 's and  $\varepsilon_i$ 's to be roughly independent.

#### 3.2 Linear constraints and the Polya posterior

Prior information involving auxiliary variables can arise in many ways. We have already discussed the case where the population mean of an auxiliary variable is known. Another case is knowing a population median. More generally one might only know that a population mean or median belongs to some interval of real numbers. Although such cases are little discussed in the usual design based literature they seem quite realistic. Another case is where a pair of auxiliary variables describe a two way table where each unit must belong to one of the cells and the population row and column totals for the numbers falling into each cell are known. Before describing the constrained Polya posterior approach to such problems we need to mention a minor technical point.

Suppose the population mean of the auxiliary variable  $x$  is known to equal  $\mu(x)$ . There will be samples where the value of  $x$  is less than  $\mu(x)$  for each unit in the sample. In such cases it would be impossible to use the constrained Polya posterior. But as a practical matter this will hardly ever happen. We will always assume that the sample we are considering is "consistent" with the prior information. This is explained in more detail just below. In our simulation studies we always reject a sample which is not consistent and select another. Again, in most cases, the probability of having to reject a sample is very small.

Each of our examples of prior information can be represented by one or more linear equality or inequality constraints. We have seen that knowing the population mean yields one linear equality constraint. If one knows that the population mean falls in some interval this yields two

linear inequality constraints. We next develop some notation that will allow us to consider a variety of situations where prior information can be described using linear equality and inequality constraints.

We assume that in addition to the characteristic of interest  $y$  the population has a set of auxiliary variables  $x^1, x^2, \dots, x^m$ . For unit  $i$  let

$$(y_i, x_i) = (y_i, x_i^1, x_i^2, \dots, x_i^m)$$

be the vector of values for  $y$  and the auxiliary variables. We assume that for any unit in the sample this vector of values is observed. We assume the prior information about the population can be expressed through a set of weighted linear equality and inequality constraints on the distinct auxiliary values in the population with weights corresponding to the proportions of the population taking these individual distinct values. We illustrate this issue more precisely by explaining how we translate this prior information about the population to the observed sample values so that we can construct pseudo-versions of the population consistent with the prior information.

Let  $s$  be a sample and, for  $j = 1, 2, \dots, n(s)$ , let  $(y_j, x_j)$  be the observed values which for simplicity we assume are distinct. Let  $p = (p_1, \dots, p_{n(s)})$  be the proportion of units which are assigned the value  $(y_j, x_j)$  in a simulated complete copy of the population. Any linear constraint on the population values of an auxiliary variable will translate in an obvious way to a linear constraint on these simulated values. For example, if the population mean of  $x^1$  is known to be less than or equal to some value, say  $b_1$ , then for the simulated population this becomes the constraint

$$\sum_{j=1}^{n(s)} p_j x_j^1 \leq b_1.$$

If the population median of  $x^2$  is known to be equal to  $b_2$  then the constraint for the simulated population becomes

$$\sum_{j=1}^{n(s)} p_j w_j = 0.5$$

where  $w_j = 1$  if  $x_j^2 \leq b_2$  and it is zero otherwise. If the population mean of  $x^2$  is less than or equal to the population mean of  $x^3$  then the simulated population constraint becomes

$$\sum_{j=1}^{n(s)} p_j (x_j^2 - x_j^3) \leq 0.$$

Hence, given a family of population constraints based on prior information and a sample we will be able to represent the corresponding constraints on the simulated  $p$  by two systems of equations

$$A_{1,s} p = b_1 \quad (2)$$

$$A_{2,s} p \leq b_2 \quad (3)$$

where  $A_{1,s}$  and  $A_{2,s}$  are  $m_1 \times n(s)$  and  $m_2 \times n(s)$  matrices and  $b_1$  and  $b_2$  are vectors of the appropriate dimensions. This generalizes the argument leading to equation 1.

We assume the sample is such that the subset of the simplex it defines by equations 2 and 3 is non-empty. For such a sample the asymptotic approximation to the constrained Polya posterior puts a uniform distribution over this subset of the simplex. Before addressing the issue of simulation from this distribution we note that it has a theoretical justification. It can be given a stepwise Bayes justification which guarantees that it will yield admissible procedures. Details are given in the appendix.

### 3.3 Computation

Let  $P$  denote the subset of the simplex which is defined by equations 2 and 3.  $P$  is a non-full dimension polytope. We would like to generate independent observations from the uniform distribution over  $P$ . Unfortunately we do not know how to do this. Instead, we use Markov chain Monte Carlo (MCMC) methods to generate dependent samples.

In particular we will use the Metropolis-Hastings algorithm which depends on using a Markov chain to generate a dependent sequence of random values for  $p \in P$ . The process works as follows. We begin by finding a starting point in  $p_0$  in the relative interior of  $P$ . This is Step 1 below. Next we choose a random direction  $d$  in  $P$ . This is a bit tricky because the dimension of  $P$  is strictly less than  $n(s) - 1$ . This is accomplished in Steps 2 and 3 below. Next we find the line segment which is the intersection of the line passing through  $p_0$  in direction  $d$  with  $P$ . This is Step 4 below. Next we choose a point at random from the uniform distribution over this line segment. This is the first observation in our Markov chain. We then repeat the process with this point playing the role of  $p_0$  to get a second random point. Letting this second random point play the role of  $p_0$  we get a third and so on. More formally our algorithm is:

- Step 1. Choose an initial positive probability vector  $p_0$  such that  $A_{1,s} p_0 = b_1$  and  $A_{2,s} p_0 < b_2$  and set  $i = 0$ .
- Step 2. Generate a random direction  $d_i$  uniformly distributed over the unit sphere in  $R^n$ .
- Step 3. Let  $d_i^*$  be the normalized projection of  $d_i$  onto the null space of  $A_{1,s}$ .
- Step 4. Find the line segment  $L_i = \{ \alpha \in R \mid p_i + \alpha d_i^* \in P \}$  and generate  $\alpha_i$  uniformly over the line segment.

Step 5. Set  $p_{i+1} = p_i + \alpha_i d_i^*$  and  $i = i + 1$  and go back to step 2.

At first glance it might not be clear what role the constraints are playing in this process. They are there however through the definition of  $P$ . The Markov chain generated in this way converges in distribution to the uniform distribution over the polytope. The convergence result of such mixing algorithms was proven by Smith (1984). If we wish to approximate the expected value of some function defined on  $P$  then the average of the function computed at the simulated values converges to its actual value. This allows one to compute point estimates of population parameters. Finding the 0.95 Bayesian credible interval approximately is more difficult.

One possibility is to run the chain for a long time; for example, we may generate 4.1 million values, throw away the first 100,000 values, and find the 0.025 and 0.975 quantiles of the remaining values. These two numbers will form our approximate 0.95 credible interval. In this manuscript we will only consider sample sizes of less than 100. For such sample sizes we have found that chains of a few million suffice.

How fast a chain mixes can depend on the constraints and the parameter being estimated. It seems to take longer to get good mixing when estimating the median than when estimating the mean. This is not surprising when one recalls that in standard bootstrap methods many more bootstrap samples are required when estimating quantiles rather than means. See for example Efron and Tibshirani (1993).

Another approach which can work well is to run the chain for a long time and then just use every  $m^{\text{th}}$  point where  $m$  is a large integer. Although this is inefficient it can give good answers when finding a 0.95 credible interval for the median.

## 4. Applications

In this section we show how various types of partial information about auxiliary variables can be incorporated in the estimation of the parameters when the constrained Polya posterior is employed. In many instances, the prior information used in the constrained Polya posterior estimation cannot be utilized by the standard frequentist methods.

### 4.1 Stratification

Stratification is a type of prior information which is commonly used in finite population sampling. We note that the usual stratified estimator can be thought of as arising from independent Polya posteriors within each stratum. Details can be found in Vardeman and Meeden (1984).

When, in addition to stratification, an auxiliary variable is present a good estimate of the population mean can be found by combining the estimates obtained from the regression estimator within each stratum. For details, see Cochran (1976). If only the population mean of the auxiliary variable is known then under standard approaches it is difficult to combine this information with stratification unless a common model is assumed across strata. The constrained Polya can incorporate both types of information which can lead to improved estimates yet it does not require the common model assumption.

To demonstrate, we constructed a stratified population of size 900 consisting of three strata. The strata sizes were 300, 200 and 400. There were two auxiliary variables, say  $x_1$  and  $x_2$ . In stratum one the  $x_{1,i}$ 's were a random sample from a gamma(10,1) distribution and the  $x_{2,i}$ 's were a random sample from a gamma(2,1) distribution. In the second stratum the  $x_{1,i}$ 's and the  $x_{2,i}$ 's were generated by the gamma(15,1) and the gamma(7,1) distributions respectively. In the third stratum the  $x_{1,i}$ 's and the  $x_{2,i}$ 's were generated by the gamma(5,1) and the gamma(3,1) distributions respectively. The characteristic of interest for the population was generated as follows:

$$\text{stratum 1: } y_i = 1 + x_{1i} x_{2i} + \varepsilon_i$$

$$\text{stratum 2: } y_i = 3 + x_{1i} + x_{1i} x_{2i} + \varepsilon_i$$

$$\text{stratum 3: } y_i = 2 + x_{2i} + x_{1i} x_{2i} + \varepsilon_i$$

where in stratum one the  $\varepsilon_i$ 's were normal(0,1), while in stratum two they were normal(0,1.5<sup>2</sup>), and in stratum three they were normal(0,3.5<sup>2</sup>). All the  $\varepsilon_i$ 's were independent.

In addition to the strata sizes we assumed that the population median of  $x_1$  and the population mean of  $x_2$  were known. We generated 500 random samples according to our sampling plan drawing 75 units such that 25 units were in the first stratum, 20 units were in the second stratum and 35 units were in the third stratum. For each sample we computed the sample mean, the usual stratified estimate which is the sum of the sample means within each stratum adjusted for the size of all strata, the constrained Polya estimate, and the corresponding 95% confidence intervals and 0.95 credible intervals for these estimates.

The results of the simulations are given in Table 1. From the table, we see that the constrained Polya estimator on average agrees with the usual stratified estimator and is essentially unbiased. But its average absolute error is much smaller than the average errors of the other two. This is to be expected since the more information an estimator uses the better it should perform and the constrained Polya estimates are using information from the auxiliary variables that is ignored by the estimates which just use stratification. Note that the constrained Polya made no assumptions about how  $y$  and  $x_1$  and  $x_2$  were related. Furthermore it is not clear



how standard methods could make use of knowing the population median of  $x_1$  and the population mean of  $x_2$ . If just information about means is available then the empirical likelihood based methods of Chen and Sitter (1999) and Zhong and Rao (2000) could be used. The results clearly show that the constrained Polya posterior is utilizing this additional information in a sensible manner.

**Table 1**  
Simulation results for the stratified example where the median of the first auxiliary variable and the mean of the second are known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	47.978	4.821	36.44	23.09	1.000
Strataest	43.395	2.072	38.22	10.35	0.942
Polyaest	43.355	1.516	40.19	6.75	0.936

In this example the constrained Polya estimates were obtained using Markov chains of length 4,000,000 after the initial 100,000 points were discarded.

4.2 Categorical auxiliary variables

Assume that the elements of a population of known size  $N$  are associated with the elements of  $k$  categorical auxiliary variables. For simplicity, we consider  $k = 2$  but the theory applies to more than two categorical variables. If one auxiliary variable takes on  $r$  distinct values and the other takes on  $c$  distinct values they allow the elements of the population to be classified into a two-way table with  $r \times c$  cells. Let  $N_{ij}$  be the number of elements in the population that belong to the  $ij$ -cell, for  $i$  in  $\{1, \dots, r\}$  and  $j$  in  $\{1, \dots, c\}$ , then  $\sum_{i=1}^r \sum_{j=1}^c N_{ij} = N$ . If the  $N_{ij}$ 's are known and  $s$  is a random sample with  $n_{ij}$  elements from the  $ij$ -cell then a good estimate of the population mean is given by

$$\sum_{i=1}^r \sum_{j=1}^c \frac{N_{ij}}{N} \bar{y}_{ij}^s,$$

where  $\bar{y}_{ij}^s$  is the mean of the  $n_{ij}$  elements from the  $ij$ -cell in the sample. This is the usual stratified estimator where the cells in the table are consider the strata.

A harder problem is the estimation of the population mean when the counts,  $N_{ij}$ 's, are not known but the marginal counts are known. Let  $N_{i\cdot} = \sum_{j=1}^c N_{ij}$  denote the marginal row counts, for  $i$  in  $\{1, \dots, r\}$  and  $N_{\cdot j} = \sum_{i=1}^r N_{ij}$  denote the marginal column counts, for  $j$  in  $\{1, \dots, c\}$ . In such cases, one way of estimating the population mean is the frequentist procedure called calibration or raking. In this procedure, given a sample  $s$ , the estimator is given by  $\sum_{k \in s} \hat{w}_k y_k$ , where the  $\hat{w}_k$ 's are not the design weights but are new weights assigned to the units in the sample. A good set of weights needs to satisfy two conditions. The first is

that the weights must preserve the known marginal counts, for example,  $\sum_{k \in s(s, j)} \hat{w}_k = N_{\cdot j}$  where  $s(\cdot, j)$  is the portion of the sample falling in the  $j^{\text{th}}$  column of the two-way table. The second is that the weights should in some sense be close to the sampling design weights  $1/\pi_k$ , where  $\pi_k = P(k \in s)$ . Depending on the function used to measure the distance different calibration estimators can be obtained. Although this is a sensible idea, selecting the right distance measure and then getting a sensible estimate of variance for the resulting estimator has no standard frequentist answer. For details, see Deville and Samdal (1992).

The Polya posterior gives an alternative approach to this problem since the information provided by the known marginal totals determines a set of linear constraints on the random weights it assigns to the units in the sample. If there are continuous auxiliary variables for which we have prior information then additional constraints can be added. To see how this could work in practice we considered a simple example with two dichotomous variables so each unit can be classified into a cell of a  $2 \times 2$  table, together with a third continuous auxiliary characteristic. We assumed four different levels of prior information.

- 1. The marginal counts for the 2-way table are known.
- 2. The marginal counts and the mean of the continuous auxiliary variable are known.
- 3. The marginal counts and the median of the continuous auxiliary variable are known.
- 4. The marginal counts are known and the mean of the continuous auxiliary variable is known to lie between two bounds. We chose the 45<sup>th</sup> and 65<sup>th</sup> quantiles of its population of values to specify these bounds.

For each case we formed a population using the following model where all the random variables are independent.

- Cell 1,1  $x_i \sim \text{gamma}(8,1)$ ,  $\varepsilon_i \sim \text{normal}(0,7^2)$  and  $y_i = 25 + 3x_i + \varepsilon_i$  for  $i$  in  $\{1, \dots, 150\}$ .
- Cell 1,2  $x_i \sim \text{gamma}(10,1)$ ,  $\varepsilon_i \sim \text{normal}(0,7^2)$  and  $y_i = 25 + 3x_i + \varepsilon_i$  for  $i$  in  $\{1, \dots, 350\}$ .
- Cell 2,1  $x_i \sim \text{gamma}(6,1)$ ,  $\varepsilon_i \sim \text{normal}(0,4^2)$  and  $y_i = 25 + 2x_i + \varepsilon_i$  for  $i$  in  $\{1, \dots, 250\}$ .
- Cell 2,2  $x_i \sim \text{gamma}(4,1)$ ,  $\varepsilon_i \sim \text{normal}(0,4^2)$  and  $y_i = 25 + 2x_i + \varepsilon_i$  for  $i$  in  $\{1, \dots, 250\}$ .

For each of the cases we generated a population and took 500 random samples of size 80 with 20 units from each cell. For each sample we computed the sample mean and the stratification estimate, assuming that the true population cell counts were known, and their corresponding 95% confidence intervals. We also computed the constrained Polya

estimates along with their 0.95 credible intervals. The constrained Polya estimates were obtained from the last 4,000,000 points of a Markov chain of size 4,100,000. The results of the simulations are given in Tables 2 through 5. The results in the tables show that the constrained Polya estimates based on known marginal counts and a known mean, median or known interval about the mean are better than the strata estimates based on known cell counts. The stratified estimates are better than the constrained Polya posterior only when the constrained Polya posterior only makes use of the known marginal counts.

**Table 2**  
Simulation results for the categorical example when just the marginal cell counts are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.805	0.919	41.107	5.396	0.976
Strataest	44.355	0.846	42.259	4.191	0.940
Polyaest	43.909	0.896	41.863	4.197	0.922

**Table 3**  
Simulation results for the categorical example when the marginal cell counts and the mean of the auxiliary variable are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.804	0.922	41.063	5.482	0.964
Strataest	44.399	0.862	42.272	4.256	0.942
Polyaest	44.506	0.510	43.257	2.497	0.960

**Table 4**  
Simulation results for the categorical example when the marginal cell counts and the mean of the auxiliary variable are assumed known

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.439	0.877	40.783	5.312	0.986
Strataest	43.927	0.884	41.804	4.244	0.940
Polyaest	43.784	0.785	42.032	3.640	0.920

**Table 5**  
Simulation results for the categorical example when the marginal cell counts are assumed known and the mean of the auxiliary variable is known to lie between its known 45<sup>th</sup> and 65<sup>th</sup> quantiles

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	43.463	0.840	40.789	5.348	0.978
Strataest	43.948	0.865	41.825	4.245	0.948
Polyaest	43.519	0.829	41.555	4.029	0.938

### 4.3 An example

In this section we consider data from the Veterans Health Administration. In 1998 the VA Upper Midwest Health Care Network administered a functional status survey of the veteran users of the VA facilities within the network (Singh, Borowsky, Nugent, Murdoch, Zhao, Nelson, Petzel and Nichol 2005). Veterans eligible for this survey were those with any outpatient encounter or inpatient stay between October 1997 and March 1998 at any one of the five VA facilities in the network. In addition to basic demographic measures, such as age and sex, the primary component of the survey was the SF36-V (Kazis, Miller, Clark, Skinner, Lee, Rogers, Spiro, Payne, Fincke, Selim and Linzer 1998). This health-related quality of life survey instrument consists of eight sub-scales of physical functioning, role limitations due to physical problems, bodily pain, general health, energy/vitality, social functioning, role limitations due to emotional problems, and mental health. These scales are combined to form physical (PCS) and mental (MCS) component summary scores. Larger scores represent better health status. VHA administrative data measuring major comorbid conditions present in the year before the survey were also collected.

From the population of one of the five facilities we selected a subpopulation comprising all of the women and a random subset of the men to form a population of 2,500 individuals. For purposes of this example the number of comorbidities was categorized into three categories to represent measures of good, average and poor health. We then selected 200 stratified random samples of size 100 from the population. The strata sizes along with the sample sizes are given in table 6. Our sampling plan over sampled the women. Such unbalanced sampling plans can often occur in practice.

**Table 6**  
The strata sizes along with the sample sizes for the Veterans Administration data

	Good	Average	Poor
F	353(20)	155(10)	117(10)
M	890(30)	493(20)	492(10)

We compared three different estimators of the mean PCS score for this population of 2,500; the sample mean which ignores the stratification, the usual stratified estimator which assume the strata sizes are known, and a constrained Polya posterior estimator which assumes that the marginal row and column totals of table 6 are known along with the average age of the individuals in the population. The population correlation between PCS and age is -0.22. The correlations of PCS with gender and with categorized comorbidity-based state of health are -0.13 and -0.28. From



the results in table 7 we see that the constrained Polya estimator performs about the same as the stratified estimator and both are a bit better than the sample mean. To compute the constrained Polya estimator we generated Markov Chains of length 7,000,000.

**Table 7**  
**Results for estimating PCS in the Veterans Administration data. The constrained Polya estimator assumes the row and column totals are known along with the average age of the individuals in the population**

Method	point estimate		95% confidence or credible intervals		
	Ave. of estimate	Ave. of absolute error	Ave. of lower bound	Ave. length	Freq. of coverage
Meanest	37.235	1.040	34.907	4.650	0.938
Strataest	36.648	0.925	34.322	4.651	0.948
Polyaest	36.644	0.925	34.344	4.605	0.958

5. Relation to empirical likelihood methods

In this section we review some frequentist methods for problems where constraints are involved and discuss their relationship to the constrained Polya posterior.

Chen and Qin (1993) considered an empirical likelihood approach to estimation in survey sampling when prior information about an auxiliary characteristic is available. To construct estimators after the sample has been observed the units in the sample are weighted to reflect the prior information. For example, suppose that the sample mean is less than the known population mean of the  $x$  values. Then positive weights, which sum to one, are selected for the sampled units such that the mean of the  $x_s$  values under the probability distribution given by the weights satisfies the known constraint. Although these weights can not be found explicitly they are easy to compute. When estimating the population mean of  $y$  the resulting estimator was first noted in Hartley and Rao (1968) and shown to be asymptotically equivalent to the regression estimator. If the population median of  $x$  is known then the units in the sample less than the known population median are given equal weights which sum to 0.5 and similarly for the sampled units with  $x$  values larger than the known population median. When estimating the population median the resulting estimator is one proposed by Kuk and Mak (1989).

An advantage of the constrained Polya posterior, and more generally of a Bayesian approach, is that it is straightforward to estimate many population quantities besides the mean without developing any new theory or methods. Given a simulated copy of the entire population which satisfies the constraints one just calculates the population parameter of interest. Then one uses such simulated values just as when one is estimating the mean.

To compare the Chen and Qin estimator of the population median of  $y$  with the constrained Polya posterior estimator when the population mean of  $x$  is known eight different populations were constructed. In half of the populations one would expect the regression estimator to do well in estimating the population mean while the remaining half did not satisfy the usual super-population model assumptions associated with the regression estimator. For each population 500 random samples of sizes 30 and 50 were taken, subject to satisfying the constraint that the sample contained values for  $x$  greater and less than the known mean. In all cases the two estimators using the prior information performed better than the sample median. These results were consistent with the simulation results of Chen and Qin. We calculated the average absolute error for the two estimators using the mean constraint. In each of the 16 different sets of simulations we then calculated the ratio of the constrained Polya posterior absolute error to that of the estimator of Chen and Qin. The range of these 16 values was 0.85 to 1.00 with a mean of 0.91. So in terms of absolute error, the constrained Polya posterior performed about 10% better, on average, than the estimator of Chen and Qin.

Suppose now that the population median of  $x$  is known. To simplify matters suppose that none of the actual values are equal to the population median of  $x$ . Let  $n_l$  be the number of units in the sample whose  $x$  values are less than the known population median of  $x$ . Then  $n_u = n(s) - n_l$  is the number of units in the sample which are on the other side of the known median. Let  $p_l = (p_1, ..., p_{n_l})$  and  $p_u = (p_1, ..., p_{n_u})$  be two probability vectors. Intuitively, a sensible posterior distribution given the sample and the known population median would be for  $p_l$  and  $p_u$  to be independent Dirichlet distributions with all parameters equal to one with each of them assigned a weight of one half so that their total sum is one. It follows from the Theorem proved in the appendix that under our sampling plan these posteriors are stepwise Bayes. Note that under these posteriors the expected values of the proportions assigned to each unit in the sample are the weights assigned to the sample by Chen and Qin. This proves the admissibility of their estimator of the population median and consequently of Kuk and Mak's. Simulation results show that this constrained Polya posterior's 0.95 credible intervals cover approximately 95% of the time except in one special case. If the sample size is small and  $y$  and  $x$  are highly correlated then the medians for the simulated populations under the constrained Polya posterior do not vary enough and the resulting intervals are too short and their coverage frequency may be considerable less than 95%.

This close relation between the empirical likelihood approach and the Polya posterior is not surprising when one



notes that in the unconstrained case the sequence of priors leading to the Polya posterior can be used to prove the admissibility of the maximum likelihood estimator for the probability vector of a multinomial distribution.

## 6. Other sampling designs

All the simulation results presented thus far have used (stratified) simple random sampling without replacement (SRS) as the sampling design. In an earlier version of the manuscript a referee wanted to know how much the behavior of estimators based on the constrained Polya posterior depended on using this design. The answer is there is some dependence but not as much as you might initially believe.

We have seen under SRS that the constrained Polya posterior (CPP) estimator behaves much like the regression estimator (REG). Formally, the regression estimator depends only on knowing the population mean of the auxiliary variable. Its properties are usually studied under simple random sampling and the estimator of its variance is only valid for large samples.

For a general design the Horvitz-Thompson estimator (HT) is often used. It is unbiased but computing the exact inclusion probabilities can be difficult. This is true, for example, if the sampling is done with selection probability proportional to the size for an auxiliary variable  $x$ , say PPS( $x$ ). In practice one simply assumes that the inclusion probability of a unit is proportional to its value of  $x$  and the resulting estimator will be approximately unbiased.

We implemented several simulation studies comparing these three methods for estimating a population total. For brevity, we present the results of two of the studies. In these studies we constructed two populations of size 500. The auxiliary variable is the same in both populations and is a random sample from a gamma distribution with shape parameter 5 and scale parameter 1. Plots of the two populations are given in figure 1. We are not suggesting that in practice one would be likely to use the regression estimator in the second population. It is presented here simply to illustrate what can happen.

For each population we took 400 random samples of size 30 and 60 under two different sampling designs. They were PPS( $x$ ) and PPS( $1/(x+5)$ ). We assumed that the population mean of  $x$  was known. For each sample we calculated the three estimates of the population total. The results from the first design are given in table 8. We see that CPP is the clear winner. The HT interval estimator's observed frequency of coverage is closest to the nominal level of 0.95. But the interval length is ridiculously long. This occurs because the reciprocals of the inclusion probabilities vary greatly. For the first population, which is

roughly linear, REG and CPP behave similarly. However, for the second population, CPP performs better than REG. It's only shortcoming here is that it under covers with the smaller sample size. Under PPS( $1/(x+5)$ ) the story is much the same although the difference between REG and CPP is much smaller for the second population. For example, when the sample size is 30 the average absolute error and frequency of coverage for REG is 131.9 and 0.875 compared to 124.3 and 0.908 for CPP. When the sample size is 60 the numbers for REG are 88.4 and 0.905 compared to 90.1 and 0.958 for CPP. The average length of their intervals are 384 and 560 respectively.

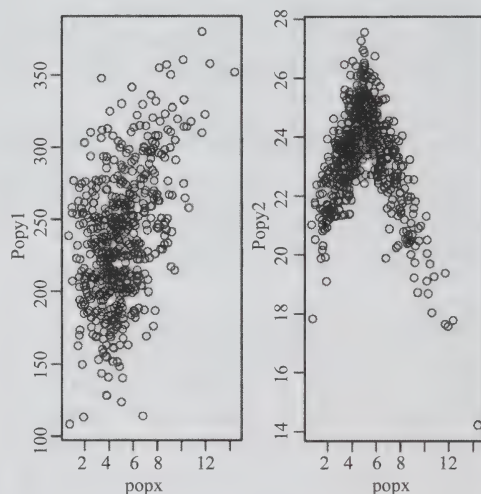


Figure 1 Plots of the two populations used in the simulations in table 8. The correlations for the two populations are 0.47 and -0.22 and their totals are 118,210 and 11,648.7

For the second population we did a second set of simulations using the PPS( $x$ ) design for sample sizes of 30 and 60. This time we assumed that the population means of  $x$  and  $x^2$  are both known. We then compared the CPP estimator which incorporates constraints on both  $x$  and  $x^2$  with the regression estimator which assumes a quadratic function of  $x$  as the model. These estimators are denoted by CPP2 and REG2 in the table. At first glance it might seem surprising that the results for CPP and CPP2 are essentially the same. But upon reflection it is what one should expect. The constrained Polya is simulating full copies of the population that are "balanced" with respect to  $x$ , that is agree with its known population mean. The additional constraint that a simulated copy of the population must be "balanced" with respect to  $x^2$  as well adds little

information. On the other hand with a sensible model the regression estimator can exploit the additional information. This results in an improved point estimator but its interval estimates still under cover.

**Table 8**  
**Simulating results for the two populations in figure 1 when estimating the population total. In each case they are based on 400 samples which were select using PPS(  $x$  ) as the design. Note abserr is the absolute value of the difference between the estimate and the true population total**

Method	Ave. of estimates	Ave. of abserr	Ave of length	Freq. of coverage
For population 1 with total = 118,210.2 for a sample size of 30				
HT	118,803.1	8,095.3	38,696.6	0.905
REG	116,838.1	3,355.3	14,136.4	0.905
CPP	117,515.7	3,277.3	14,330.7	0.905
for a sample size of 60				
HT	119,139.2	5,395.6	28,233.3	0.952
REG	117,041.4	2,213.2	9,561.0	0.910
CPP	118,041.4	2,195.3	11,836.5	0.938
For population 2 with total = 11,648.7 for a sample size of 30				
HT	11,737.2	783.5	4,012.0	0.945
REG	11,800.3	179.7	533.0	0.745
CPP	11,689.9	122.4	535.4	0.900
REG2	11,660.0	97.2	382.3	0.862
CPP2	11,689.9	122.4	535.4	0.900
CPPbd	11,683.2	116.5	537.0	0.918
for a sample size of 60				
HT	11,774.2	564.8	2,908.2	0.955
REG	11,795.8	155.2	373.1	0.635
CPP	11,647.9	80.4	524.4	0.978
REG2	11,663.1	66.7	266.2	0.895
CPP2	11,651.2	88.4	523.6	0.962
CPPbd	11,644.6	83.9	552.1	0.978

For the second population we did a third set of simulations using PPS(  $x$  ) as the design for sample sizes of 30 and 60. In this case we assumed that the population mean of  $x$  was contained in the interval (4.45, 5.53). These are the 0.45 and 0.65 quantiles of the  $x$  population. The mean of this population is 5.02. The results are in table 8 under the label CPPbd. We see that the results are very similar to those where the population mean of  $x$  was assumed to be known.

All three estimators are using the information contained in the auxiliary variable  $x$  but the HT estimator is the only one that depends on knowing the sampling design. As we have noted, it is well known that Bayesian estimators do not use the design probabilities in their computation. In these examples we see that CPP is making effective use of the information contained in the auxiliary variable. In general, the Polya posterior and variations on the Polya posterior,

like the Constrained Polya posterior, do not rely directly upon simple random sampling, stratified random sampling, or any other design. Their suitability and their performance are dependent upon the agreement of the structure underlying the population and the structure specified in the chosen predictive distribution.

The basic idea underlying the CPP is that one should use the sample and the available auxiliary information to simulate complete representative copies of the population. In simple examples like those given above we see that its point estimator should have excellent frequentist properties for a wide class of designs and the performance of its interval estimator will be adequate if the sample size is not too small. Does this mean that it can automatically adjust “bad” samples to get good estimates? Not really since with a very bad sample, one that agrees poorly with the known prior information, two bad things can happen. First, extremely unbalance or biased samples will introduce some bias into the point estimate. Second, they will severely constrain the possible values of  $p$  under the CPP and result in a posterior variance that is too small, which will lead to interval estimates that are too short and under cover the quantity of interest. In more complicated situations further study needs to be done to discover when the CPP can profitably be employed.

### 7. Final remarks

One problem with standard frequentist methods is that each different problem demands its own solution. Estimating the population median of  $y$  when the population mean of  $x$  is known is a different problem than estimating the mean of  $y$  when the mean of  $x$  is known. Also, if the population mean of  $x$  is not known exactly but is only known to belong to some interval of values then the standard frequentist methods cannot make use of this information. A strength of a Bayesian approach is that once you have a posterior distribution which sensibly combines the sample with the prior information inference can be done for many population parameters of interest simply by simulating completed copies of the population.

Here we have argued that the constrained Polya posterior is a sensible method of introducing objective prior information about auxiliary variables into a noninformative Bayesian approach to finite population sampling. The resulting point estimators have a stepwise Bayes justification which guarantees their admissibility. Their 0.95 credible intervals will usually be approximate 95% confidence intervals and they give sensible answers for problems where there are no standard frequentist procedures available. This demonstrates an important strength of the Polya posterior. Once you can simulate sensible copies of the

entire population inference for a variety of problems becomes straightforward. On the downside, one needs to use MCMC methods for their calculation. All our computations were done in R (R Development Core Team 2005). Two of the authors have recently released an R package *polyapost* which makes it easy for others to use our methods. Here we have restricted ourselves to samples of less than 100. This was just a matter of convenience so we could do our simulations in a reasonable amount of time. In practice for a larger specific sample one just needs to run a longer chain. Then one can use some of the standard diagnostics to decide whether or not it seems to have converged.

## Appendix

### An admissibility proof

The basic theoretical justification for point estimators arising from the Polya posterior is that they are admissible. The proofs of admissibility use the stepwise Bayes nature of the Polya posterior. This section presents a proof for point estimators based on the constrained Polya posterior.

In these stepwise Bayes arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order is important. A different prior distribution is defined on each of the subsets. Then, the Bayes procedure is found for each sample point that receives positive probability under the first prior. Next, the Bayes procedure is found for each sample point which receives positive probability under the second prior and which was not considered under the first prior. Then, the third prior is considered and so on. For a particular sample point the value of the stepwise Bayes estimate is the value of the Bayes procedure from the step at which it was considered. It is the stepwise Bayes nature of the Polya posterior that explains its somewhat paradoxical properties. Given a sample, it behaves just like a proper Bayesian posterior but one never has to explicitly specify a prior distribution. For more details and discussion on these points see Ghosh and Meeden (1997).

To prove the admissibility of the estimators arising from the Polya posterior for the parameter space  $[0, \infty)^N$  the main part of the stepwise Bayes argument first assumes that the parameter space is  $\Lambda^N$ , where  $\Lambda$  is an arbitrary finite set of positive real numbers. Once admissibility has been demonstrated for such general  $\Lambda$ , admissibility for the parameter space  $[0, \infty)^N$  follows easily. A similar argument will be used for the constrained Polya posterior.

Dealing with constraints on finite populations introduces some technical problems which are difficult to handle. For this reason, we will suppose that the population is large enough compared to the sample size that the approximate

form of the Polya posterior involving the Dirichlet distribution is appropriate. For simplicity we assume that the population  $U$  is infinite.

We assume that for all  $j$  in  $U$ ,  $(y_j, X_j) = a_i$  for some  $i$  in  $\{1, \dots, k\}$ , where  $a_i = (a_{i1}, \dots, a_{i(m+1)})$  are distinct vectors in  $R^{m+1}$  and where  $k$  can be very large. That is, the vectors  $(y_j, X_j)$  can take on only a finite number of values. If  $p_i$  is the proportion of  $(y_j, X_j)$ 's in the population which are equal to  $a_i$ , for  $i$  in  $\{1, \dots, k\}$ , then the population mean of  $Y$  is  $\sum_{i=1}^k p_i a_{i1}$ .

We assume that there is prior information available about the auxiliary variables  $X^i := \{x_j^i \mid j \in U\}$  for  $i$  in  $\{1, \dots, m\}$ , which gives rise to linear equalities and inequalities involving the proportions  $p$  of the form

$$A_1 p = b_1 \quad (4)$$

$$A_2 p \leq b_2 \quad (5)$$

where  $A_1, A_2$  are  $m_1 \times k$  and  $m_2 \times k$  matrices and  $b_1, b_2$  vectors of appropriate dimensions. In this setting, for instance, we may want to estimate

$$\mu(p) = \sum_{i=1}^k p_i a_{i1}$$

subject to the constraints in equations 4 and 5 and where  $\sum_{i=1}^k p_i = 1$  with  $p_i \geq 0$ , for all  $i$  in  $\{1, \dots, k\}$ .

Consider a sample  $s$  of size  $n$  which for notational convenience we will assume consists of  $n$  distinct  $a_i$ 's. Let  $a_s$  denote this set of values. We then let  $A_{1,s}$  and  $A_{2,s}$  be the  $m_1 \times n$  and  $m_2 \times n$  matrices which are just  $A_1$  and  $A_2$  restricted to the columns corresponding to the members of  $a_s$ . Let  $p_s$  be  $p$  restricted to the members of  $a_s$ . Then the constraints on the population given in equations 4 and 5 translate into the following constraints

$$A_{1,s} p_s = b_1 \quad (6)$$

$$A_{2,s} p_s \leq b_2 \quad (7)$$

for the random weights assigned to members of the sample. That is, given a sample the constrained Polya posterior is just the uniform distribution over the subset of the simplex defined by equations 6 and 7.

A technical difficulty when proving admissibility under constraints is that even when the population satisfies the stated constraints it is always possible to get a sample which fails to satisfy them. There are several ways one can handle such cases. One possibility is to assume that the constraints are wrong and just ignore them. This tactic was used in Nelson and Meeden (1998). Another possibility is to use prior information to augment the sample so that it satisfies the constraints. This can be messy and your answer can depend strongly on how you adjust the sample. We will take a third approach here.



We will assume the sampling design is simple random sampling and that our prior information must be correct. In such a situation it might make sense to reject any sample which does not satisfy the constraints since it is clearly an unrepresentative sample. More specifically, suppose we take a simple random sample of size  $n$  from the population and observe all  $x_j$ 's in the sample. Let  $p^s = (p_1^s, \dots, p_k^s)$  be the proportions of the possible vectors for the  $x_j$  that are observed in the sample. The element  $p_i^s$  is zero whenever the vector  $a_i = (a_{i2}, \dots, a_{i(m+1)})$  does not appear in the sample. If  $p^s$  satisfies equations 6 and 7 we keep the sample, if not we discard it and try again. We will call this sampling plan constraint restricted random sampling. In practice, for typical constraints, it will almost never be necessary to discard a sample. Although this is a sampling plan that would never be used it is not a bad approximation to what is actually done.

More formally, let  $Z_i$  be the number of  $(y_j, X_j)$ 's in the sample that equal  $a_i$ , for  $i$  in  $\{1, \dots, k\}$ , then  $(Z_1, \dots, Z_k)$  is Multinomial( $n, p_1, \dots, p_k$ ) where the parameter values belong to

$$P := \left\{ (p_1, \dots, p_k) \mid A_1 p = b_1, A_2 p \leq b_2, \sum_{i=1}^k p_i = 1, \right. \\ \left. \text{and } p_i \geq 0 \forall i \in \{1, \dots, k\} \right\}. \quad (8)$$

For a given sample  $s = (z_1, \dots, z_k)$  let

$$P^s := \{p \mid p \in P \text{ and } p_i = 0 \text{ whenever } p_i^s = 0 \text{ for } i=1, \dots, k\}. \quad (9)$$

We see that we keep a sample if and only if  $P^s$  is not empty.

Denote the  $k-1$  dimensional simplex by

$$F := \left\{ (p_1, \dots, p_k) \mid \sum_{i=1}^k p_i = 1, p_i \geq 0 \forall i \in \{1, \dots, k\} \right\}.$$

For  $i = 1, \dots, k$  let  $e_i$  denote the vertices of  $F$ . The  $e_i$ 's are the unit vectors whose  $i^{\text{th}}$  value is 1 and is 0 elsewhere.

Now  $P$  is a convex polytope which is the intersection of  $F$  with the space

$$G := \{(p_1, \dots, p_k) \mid A_1 p = b_1, A_2 p \leq b_2\}.$$

A partition of the parameter space  $P$  can be found in the following way. Let  $F_j$  denote the set of faces of dimension  $j$  of the simplex  $F$ ,  $j = 0, 1, \dots, k-1$ . Then  $F_0$  is the set of its vertices,  $F_j$  is the collection of the convex hulls of all combinations of  $j+1$  vertices, for  $j = 1, \dots, k-2$  and  $F_{k-1}$  is the simplex  $F$ . If  $\text{int}(F_j)$  is the set of the interiors of the faces of dimension  $j$ , for  $j = 1, \dots, k-1$ , then  $\{F_0, \text{int}(F_1), \dots, \text{int}(F_{k-1})\}$  determines a partition of the simplex  $F$ . If  $G_0 := F_0 \cap G$  and  $G_j := \text{int}(F_j) \cap G$  for  $j = 1, \dots, k-1$  then  $\{G_0, G_1, \dots, G_{k-1}\}$  is a partition of the parameter space  $P$ . Note that some of  $G_j$ 's might be

empty. The stages of the stepwise Bayes argument follow the nonempty members of the  $G_j$ 's.

If  $Z$  is the sample space of the counts  $(Z_1, \dots, Z_k)$  then for  $p \in P$  the distribution of the counts, say  $f_p(z \mid p)$  is Multinomial( $n, p_1, \dots, p_k$ ) when the sample size is  $n$ . Let  $P_F$  be the restriction of the parameter space  $P$  to  $F$ , where  $F$  is any subset of  $P$  and  $Z_{P_F}$  be the restriction of the sample space  $Z$  determined by  $P_F$ .

We are now ready to prove the admissibility of the constrained Polya posterior estimator of  $\mu(p)$  over  $P$ . Suppose we are at the stage where we are considering  $G_j$ , for some  $j \geq 0$ . Assume  $G_j = G \cap F$  for some subset is nonempty. There are two possible cases.

Case 1. If the dimension of  $G_j$  is zero, i.e., it consists of one vector, say  $p^0$ , then we take the prior that puts unit mass on this vector. The posterior also then puts unit mass on this vector and if  $z$  is the unique member of  $Z_{P_F}$  then the Bayes estimator is  $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z) = \mu(p^0)$ .

Case 2. If the dimension of  $G_j$  is greater than zero then the distribution of  $(Z_1, \dots, Z_k)$  restricted to  $Z_{P_F}$  is

$$f_{P_F}(z \mid p) = \frac{f_p(z \mid p)}{\sum_{z \in Z_{P_F}} f_p(z \mid p)}.$$

The prior we consider on  $P_F$  is

$$\pi_{P_F}(p) \frac{\sum_{z \in Z_{P_F}} f_p(z \mid p)}{\prod_{\{i \mid p \in P_F, p_i > 0\}} p_i},$$

which can be normalized to be a proper prior since  $\sum_{z \in Z_{P_F}} f_p(z \mid p)$  can be written as  $g(p) \prod_{\{i \mid p \in P_F, p_i > 0\}} p_i$  where  $g(p)$  is a bounded function of  $p$ . With this prior, the posterior distribution is the Dirichlet density kernel restricted to  $P_F$ ,

$$f_{P_F}(p \mid z) \propto f_p(z \mid p) \pi_{P_F}(p) \propto \prod_{\{i \mid p \in P_F, p_i > 0\}} p_i^{z_i - 1}.$$

The Bayes estimator of  $\mu(p)$  against  $\pi_{P_F}$ , where  $p$  belongs to  $P_F$ , is  $\delta_{\pi_{P_F}}(z) = E(\mu(p) \mid z)$  for all  $z$  in  $Z_{P_F}$ . Hence, if we use the sequence of priors

$$\{\{\pi_{P_F \mid F \in G_0}\}, \{\pi_{P_F \mid F \in G_1}\}, \{\pi_{P_F \mid F \in G_2}\}, \dots, \{\pi_{P_F \mid F \in G_{k-1}}\}\},$$

ignoring the empty sets at each step, then the estimator  $\delta(z)$  defined by

$$\delta(z) = \delta_{\pi_{P_F}}(z) \text{ for } z \in Z_{P_F}, F \in G_i, i = 1, \dots, \gamma - 1, \quad (10)$$

where  $\gamma = k$  if  $k < n$  and  $n$  if  $k \geq n$ , is an admissible estimator for  $\mu(p)$ . This concludes the proof of the following theorem.

**Theorem 1.** Under the constraint restricted random sampling plan defined by equations 2 and 3 with parameter

space defined in equation 8 the constrained Polya posterior estimator given in equation 10 for estimating the population mean is stepwise Bayes and hence admissible under squared error loss.

### Acknowledgements

The research of Glen Meeden and Radu Lazar was supported in part by NSF Grant DMS 0406169. The authors would like to thank the referees and associate editor for several helpful comments.

### References

- Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Series B*, 44, 388-393.
- Booth, J.G., Bulter, R.W. and Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.
- Chambers, R.L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., and Sitter, S.S. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- Cochran, W. (1976). *Sampling Techniques*. New York: John Wiley & Sons, Inc., 3<sup>rd</sup> Edition.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Efron, B., and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.
- Feller, W. (1968). *An Introduction of Probability Theory and its Applications*. New York: John Wiley & Sons, Inc., Volume I.
- Ghosh, J.K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays*. By (Ed. D. Basu). New York: Springer-Verlag.
- Ghosh, M., and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. London: Chapman and Hall.
- Gross, S. (1980). Median estimation in survey sampling. In *Proceedings of the Section of Section on Survey Research Methods*, American Statistical Association, 181-184.
- Hartley, H.O., and Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.
- Kazis, L.E., Miller, D.R., Clark, J., Skinner, K., Lee, A., Rogers, W., Spiro, A. 3<sup>rd</sup>, Payne, S., Fincke, G., Selim, A. and Linzer, M. (1998). Health related quality of life in patients served by the department of veterans affairs. *Archives of Internal Medicine*, 158, 626-632.
- Kuk, A.Y.C., and Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, 51, 261-269.
- Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.
- Mak, T.K., and Kuk, A. (1993). A new method for estimating finite population quantiles using auxiliary information. *Canadian Journal of Statistics*, 21, 29-38.
- Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sankhyā, Series A*, 61, 133-144.
- Meeden, G. (2005). A noninformative bayesian approach to domain estimation. *Journal of Statistical Planning and Inference*, 129, 85-92.
- Nelson, D., and Meeden, G. (1998). Using prior information about population quantiles in finite population sampling. *Sankhyā, Series A*, 60, 426-445.
- Nelson, D., and Meeden, G. (2006). Noninformative nonparametric finite population quantile estimation. *Journal of Statistical planning and Inference*, 136, 53-67.
- R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- Singh, J.A., Borowsky, S.J., Nugent, S., Murdoch, M., Zhao, Y., Nelson, D., Petzel, R. and Nichol, K.L. (2005). Health related quality of life, functional impairment, and health care utilization by veterans: Veterans' quality of life study. *Journal of American Geriatric Society*, 53, 108-113.
- Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32, 1296-1308.
- Valliant, R., Dorfman, A. and Royall, R. (2000). *Finite Population Sampling and Inference, a Predictive Approach*. New York: John Wiley & Sons, Inc.
- Vardeman, S., and Meeden, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.
- Woodruff, R.S. (1952). Confidence intervals for the median and other positive measures. *Journal of the American Statistical Association*, 47, 635-646.
- Zhong, B., and Rao, J.N.K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika*, 87, 920-938.

# Optimal sample allocation for design-consistent regression in a cancer services survey when design variables are known for aggregates

Alan M. Zaslavsky, Hui Zheng and John Adams<sup>1</sup>

## Abstract

We consider optimal sampling rates in element-sampling designs when the anticipated analysis is survey-weighted linear regression and the estimands of interest are linear combinations of regression coefficients from one or more models. Methods are first developed assuming that exact design information is available in the sampling frame and then generalized to situations in which some design variables are available only as aggregates for groups of potential subjects, or from inaccurate or old data. We also consider design for estimation of combinations of coefficients from more than one model. A further generalization allows for flexible combinations of coefficients chosen to improve estimation of one effect while controlling for another. Potential applications include estimation of means for several sets of overlapping domains, or improving estimates for subpopulations such as minority races by disproportionate sampling of geographic areas. In the motivating problem of designing a survey on care received by cancer patients (the CanCORS study), potential design information included block-level census data on race/ethnicity and poverty as well as individual-level data. In one study site, an unequal-probability sampling design using the subjects' residential addresses and census data would have reduced the variance of the estimator of an income effect by 25%, or by 38% if the subjects' races were also known. With flexible weighting of the income contrasts by race, the variance of the estimator would be reduced by 26% using residential addresses alone and by 52% using addresses and races. Our methods would be useful in studies in which geographic oversampling by race-ethnicity or socioeconomic characteristics is considered, or in any study in which characteristics available in sampling frames are measured with error.

**Key Words:** Descriptive population quantity; Measurement error; Neyman allocation; Regression models; Sample designs; Surveys.

## 1. Introduction

A sample survey is to be designed to obtain data that will be used to estimate coefficients of one or more regression models. Information about the population distribution of the covariates is available, and also some covariate information is available in the sampling frame. How can this information be used to make the survey design more efficient? How much can variance be reduced with such a design, relative to simple random sampling, and how is that answer affected if the frame only provides covariate distributions aggregated over groups, but not for individual subjects?

These questions were motivated by design of a survey of health care processes (such as provision of chemotherapy when appropriate) and outcomes (such as quality of life after treatment) for a large sample of cancer patients at seven sites in the United States, conducted as part of the CanCORS (Cancer Care Outcomes Research and Surveillance) study (Ayanian, Chrischilles, Wallace, Fletcher, Fouad, Kiefe, Harrington, Weeks, Kahn, Malin, Lipscomb, Potosky, Provencale, Sandler, Vanryn and West 2004). Among the primary objectives of this study was to estimate joint effects of race and income on these measures, using regression models that include both of these patient characteristics. However, only limited data were available

when patients were sampled for enrollment in the study. Prior experience suggested that race and residential address might be determined with reasonable accuracy at the time cases were ascertained for possible study recruitment, but income could not be determined until the subject was recruited and interviewed, and could not practically be collected in a screening interview. We undertook the research reported here to determine how the available patient data could be combined with census data on race-income distributions in census blocks to sample patients disproportionately and thereby improve estimates of race and income effects.

Such concerns arise frequently when survey data will be used to estimate coefficients of one or more regression models. For example, the National Health Interview Survey (NHIS) uses geographical oversampling together with a screening interview to oversample Black and Hispanic respondents for improved domain estimation (Botman, Moore, Moriarity and Parsons 2000, page 12); NHIS data have been used extensively in regression analyses, of which domain estimation is a special case. Sastry, Ghosh-Dastidar, Adams and Pebley (2005, pages 1013-1014) oversampled census tracts by minority composition, using simulations to evaluate the power of various designs for regression analyses of interest. The Youth Risk Behavior Surveillance

1. Alan M. Zaslavsky, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.; Hui Zheng, Biostatistics Center, Massachusetts General Hospital, 50 Staniford Street, Boston, MA 02114, U.S.A.; John Adams, RAND Corporation, 1776 Main Street, Santa Monica, CA 90401.



System oversamples *schools* in high-minority PSUs to improve precision of estimates for minority racial/ethnic groups (Eaton, Kann, Kinchen, Ross, Hawkins, Harris, Lowry, McManus, Chyen, Shanklin, Lim, Grunbaum and Wechsler 2006, pages 2-3).

The literature on optimal design of experiments is extensive. Design objectives for surveys, however, differ in important ways from those for experiments, in which the researcher can arbitrarily assign *a priori* identical units to treatments. A strongly model-based approach to estimation of regression coefficients would suggest selection of a suitable set of high-leverage observations, much as in design of an experiment (Royall 1970), but the application of these principles to survey design is controversial (see Sec. 4). The design-based approach requires a sample that is representative, through a known probability mechanism, of a defined population; intermediate positions are also possible (Sec. 2.5, 3.4). From this perspective, the sampler is not free to select, for example, 100 white respondents from a convenient primarily white neighborhood and 100 Hispanic respondents from a convenient primarily Hispanic neighborhood and call the sample “representative” for estimating differences between whites and Hispanics. The objects of design-based inference are quantities that describe the population; in the case of regression we will refer to “descriptive population quantity” or DPQ regressions (Pfeffermann 1993, pages 319-321).

Since Neyman (1934), the extensive literature on optimal design of surveys (reviewed in standard texts like Cochran 1977 or Särndal, Swensson and Wretman 1992) has primarily focused on estimation of simple quantities such as a mean or ratio, or of several such quantities (Kish 1974; Bellhouse 1984; Chromy 1987). Although variance estimation for design-based estimates of regression coefficients has received considerable attention (Fuller 1975; Fuller 1984; Binder 1981; Binder 1983), relatively little attention has been given to the corresponding optimal sample designs. (Regression-assisted estimation of a mean (Cassel, Särndal and Wretman 1976; Särndal, Swensson and Wretman 1992, Sec. 12.2) is a distinct problem.)

Furthermore, characteristics that might be used to define an unequal-probability sampling scheme are likely to be recorded with error in sampling frames, because they are based on aggregated data or because the characteristics associated with a unit (such as an address or a household) change over time. Such errors can greatly affect the efficiency of a putatively optimal sampling scheme; see Morris, Newhouse and Archibald (1979, Sec. III) on stratified sampling for domain estimation and Thomsen, Tesfu and Binder (1986) on probability-proportional-to-size sampling. Waksberg (1973, 1995) considers stratification by census blocks on a single aggregated characteristic for

estimation of means for domains such as racial/ethnic groups or the poor, with or without a subsequent screening interview.

Our objective in this article is to describe optimal designs for samples that will be used in DPQ (design-weighted) regression analysis, in the sense of minimizing the weighted sum of variances of some preselected linear combinations of regression coefficients. We also consider some classes of estimands and corresponding estimators that depart from the DPQ approach to improve efficiency. In Section 2, we establish notation and derive optimal sampling rates for DPQ regression under scenarios representative of the individual and area-level information that might be encountered in population surveys with imperfect frames. We first assume that exact design information is available in the sampling frame and then generalize to situations in which some variables are available only as aggregates for subdomains or from inaccurate data. We next consider optimal estimation of combinations of coefficients from more than one model and of flexible combinations of coefficients. In Section 3 we estimate the potential benefits of these methods for a survey in the CanCORS study sites, using block-level census data on race/ethnicity and poverty. Finally, in Section 4 we consider the relevance of the DPQ approach and possible extensions of the methodology.

## 2. Optimal design calculations

### 2.1 Notation

Suppose that the target population is divided into cells indexed by  $b = 1, 2, \dots, B$ , with elements indexed by  $k = 1, 2, \dots, K_b$  in cell  $b$ . With each element is associated a covariate vector  $\mathbf{x}_{bk}$  with  $\mathbf{x}_{bk} = (\mathbf{u}_{bk}', \mathbf{t}_{bk}')$ , where  $\mathbf{u}_{bk}$  is the component observed for identifiable individuals. The distribution of  $\mathbf{t}_{bk}$  in each cell is known but the values for individuals are not observed; thus the cell is the unit of aggregation for some or all of the design variables. Hence we know the finite population values  $\mathbf{T}_b = (\mathbf{t}_{b1}, \mathbf{t}_{b2}, \dots, \mathbf{t}_{bK_b})'$  but cannot identify the rows with individuals. Define  $\bar{\mathbf{t}}_b = \mathbf{1}' \mathbf{T}_b / K_b$ , the mean of  $\mathbf{t}$  in cell  $b$ .

Associated with sampling each element is a cost  $c_{bk}$ . A sampling plan is defined by assigning a probability of selection  $\pi_{bk}$  to each element. Assume a constraint on expected cost,

$$\sum_{b,k} c_{bk} \pi_{bk} \leq C. \quad (1)$$

To simplify the presentation, we also assume that the sampling rate is low and potential benefits of stratification are minimal, so the design can be described approximately as unstratified unequal-probability sampling with replacement. We also assume single-stage element sampling. The

population is  $U = \{(b, k): b = 1, 2, \dots, B; k = 1, 2, \dots, K_b\}$  and a sample is  $S \subset U$ .

The population-descriptive ordinary least squares (OLS) regression coefficient, corresponding to the model  $y_{bk} = \beta' \mathbf{x}_{bk} + \varepsilon_{bk}$  with  $\varepsilon_{bk} \sim [0, \sigma^2]$ , is  $\beta_U = (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{X}'_U \mathbf{y}_U$ , where subscript  $U$  signifies matrices or vectors corresponding to the entire population. (Here  $[0, \sigma^2]$  signifies a distribution with mean 0 and variance  $\sigma^2$ , but unspecified form.) Then

$$\hat{\beta} = (\mathbf{X}'_S \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}'_S \mathbf{W}_S \mathbf{y}_S \quad (2)$$

is the usual design-based estimator of  $\beta$ , where  $S$  signifies that only the rows corresponding to the sample are included, and  $\mathbf{W}$  is the diagonal matrix of weights  $1/\pi_{bk}$ .

To design the survey, we must make some assumptions about the distribution of outcomes  $y_{bk}$ , even if we would not rely on the same assumptions in analysis of the data. Specifically, we assume that the outcomes are generated by a model  $\xi: y_{bk} = \mathbf{x}'_{bk} \beta + \varepsilon_{bk}$ , with independent  $\varepsilon_{bk} \sim [0, \sigma^2_{bk}]$  and known  $\sigma^2_{bk}$  (up to a constant factor). Note that the distributions of the design variables  $\mathbf{x}_{bk}$  and the residuals are relevant to optimization of the design, but the value of  $\beta$  is not since it does not affect the variance of the regression estimators. Furthermore, the assumption of independent residuals from a regression model might be more reasonable than independence of data values. We allow for heteroscedasticity, even when fitting an OLS model. OLS coefficients (including special cases such as the overall mean or domain means) are often useful descriptive statistics even if the OLS model does not actually hold, but if information about heteroscedasticity is available it can be used to make the design more efficient.

## 2.2 Optimal DPQ regression design with individual-level variables only

Consider first the case in which  $\mathbf{t}$  is empty, so  $\mathbf{x}_{bk} = \mathbf{u}_{bk}$ , reflecting a scenario in which all relevant design variables (race and income in our CanCORS design) are available to the researcher before sampling. Since the cells now consist of single cases we drop the subscript  $b$ , writing  $\hat{\beta} = (\sum_S \mathbf{w}_k \mathbf{x}_k \mathbf{x}'_k)^{-1} (\sum_S \mathbf{w}_k \mathbf{x}_k y_k)$ . Then for any fixed linear combination of coefficients with weights  $\mathbf{a}$ , assuming that the first factor is a design-consistent estimator (after scaling) of  $N(\mathbf{X}'_U \mathbf{X}_U)^{-1}$ , we have the expectation under sampling of the model-based variance (White 1980) of the estimator,

$$\begin{aligned} V_a &= E_\pi \text{Var}_\xi \mathbf{a}' \hat{\beta} \\ &\approx \mathbf{a}' (\mathbf{X}'_U \mathbf{X}_U)^{-1} \left( E_\pi \text{Var}_\xi \sum_{k \in S} \mathbf{x}_k y_k / \pi_k \right) (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{a} \\ &= \mathbf{a}' (\mathbf{X}'_U \mathbf{X}_U)^{-1} \left( \sum_{k \in U} (\sigma_k^2 / \pi_k) \mathbf{x}_k \mathbf{x}'_k \right) (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{a}. \quad (3) \end{aligned}$$

For design-based inference, the relevant measure is the average variance under the sampling design over possible populations obtained under the model  $\xi$ ,  $E_\xi \text{Var}_\pi \mathbf{a}' \hat{\beta}$  (the “anticipated variance” of Isaki and Fuller 1982; see also Bellhouse 1984, sec. 1); this quantity is approximately equal to the expected model-based variance (see Appendix for proof and asymptotic conditions).

By the typical Lagrange multiplier argument for optimal allocation problems (e.g., Valliant, Dorfman and Royall 2000, pages 169-170),  $V_a$  is minimized subject to the expected cost constraint (1) when  $\partial V_a / \partial \pi_k = c_k \lambda$  for some constant  $\lambda$  and all  $k$ , so  $\pi_k \propto \sigma_k (\mathbf{a}' \mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{x}_k) / \sqrt{c_k}$ . Thus the optimal sampling rate is higher for cases with greater model variance and lower cost (as in the usual case of estimation of a mean) and also for cases with greater leverage in the regression. This result differs from the standard model-based calculations for optimal experimental design, which would allocate the entire sample to a few high-leverage design points. The design-consistent estimator of the DPQ regression does not assume the correctness of the model and therefore requires that every case have a positive probability of selection. Thus, for estimation of a ratio  $\beta$  under a homoscedastic model  $y_k = \beta x_k + \varepsilon_k$ , model-based estimation would suggest selection of the units with the largest values of  $x_k$ , but our probabilities of selection are proportional to  $x_k$ .

Typically, more than one estimand will be of interest in a study; CanCORS is intended to estimate both race and income effects. We generalize (3) to simultaneous estimation of several linear combinations of coefficients by optimizing a weighted sum of variances  $V = \sum_i d_i V_{\mathbf{a}_i}$ , where  $i$  indexes the estimands. By the same arguments the optimal sampling probabilities for this objective are

$$\pi_k \propto \sigma_k \left( \sum_i d_i (\mathbf{a}'_i (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{x}_k)^2 / c_k \right)^{1/2}. \quad (4)$$

With some choices of the  $\{\mathbf{a}_i\}$ , strict adherence to (4) could lead to arbitrarily small  $\pi_{bk}$  (and hence arbitrarily large weights) for cases with leverage approaching zero. To prevent this, we could set a positive floor on the  $\pi_k$ . Alternatively, by making estimation of the population mean one of the objectives (Section 2.4), we guarantee sampling with positive probability over the entire population. Either method makes the design more robust against error in the approximate calculation of leverage and better prepared for possible post hoc decisions to estimate quantities not foreseen in the original design plan (Section 2.6). Furthermore, reasonably good estimation of means is needed to guarantee design-consistency of the first factor of (2).



### 2.3 Optimal design with individual- and aggregate-level variables

Now suppose that the covariate vector  $\mathbf{t}_{bk}$  is nonempty and  $\mathbf{u}_{bk}$  is constant in each cell, as when aggregated design information is available for cells corresponding to covariate classes of  $\mathbf{u}$  within blocks. In CanCORS, if race ( $\mathbf{u}$ ) but not income ( $\mathbf{t}$ ) is known for individual subjects, and income distributions are available for each race in each census block, we would define cells to consist of people of a single race in a single census block.

Since cases in the same cell cannot be distinguished on covariates we further assume that  $\sigma_{bk} = \sigma_b$  and  $c_{bk} = c_b$  are *a priori* constant across the cell, so the optimal design also makes  $\pi_{bk} = \pi_b$  constant in each cell.

We can now rewrite (3) as

$$V_a \approx \mathbf{a}'(\mathbf{X}'_U \mathbf{X}_U)^{-1} \left( E_\pi \text{Var}_\zeta \sum_S \mathbf{x}_{bk} y_{bk} / \pi_b \right) (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{a} \\ = \mathbf{a}'(\mathbf{X}'_U \mathbf{X}_U)^{-1} \left( \sum_{b,k} (\sigma_b^2 / \pi_b) \mathbf{S}_b \right) (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{a}, \quad (5)$$

where

$$\mathbf{S}_b = \begin{pmatrix} \mathbf{u}_b \mathbf{u}_b' & \mathbf{u}_b \bar{\mathbf{t}}_b' \\ \bar{\mathbf{t}}_b \mathbf{u}_b' & \mathbf{S}_{Tb} \end{pmatrix}$$

is the matrix of mean squares and crossproducts in cell  $b$ , with  $\mathbf{S}_{Tb} = \mathbf{T}_b' \mathbf{T}_b / K_b$ . The optimal sampling probabilities corresponding to (4) are then

$$\pi_b \propto \sigma_b \left( \sum_i d_i \mathbf{a}_i' (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{S}_b (\mathbf{X}'_U \mathbf{X}_U)^{-1} \mathbf{a}_i / K_b c_b \right)^{1/2}. \quad (6)$$

If  $\mathbf{t}$  is measured through a census of each cell, then  $\bar{\mathbf{t}}_b$  and  $\mathbf{S}_{Tb}$  are known exactly. The same principles apply, however, if  $\mathbf{S}_{Tb}$  is not directly observed but instead is estimated under a model  $\zeta$ . We then replace  $\bar{\mathbf{t}}_b$  and  $\mathbf{S}_b$  in (5) with predictive expectations  $\tilde{\mathbf{t}}_b = E_\zeta \bar{\mathbf{t}}_b$  and  $\tilde{\mathbf{S}}_b = E_\zeta \mathbf{S}_b$ . Examples might include the following situations: (1) data for each cell are only available for a sample, (2) design data are old and the distribution of design variables in the cell may have changed over time, or (3) data on individual elements are measured with error. Similarly, the distribution of  $\mathbf{t}$  might be available only for a supercell that contains multiple values of  $\mathbf{u}$  (for example, race and census block of residence are known for each individual, but the income distribution is known for the block as a whole but not for each race within the block), so  $\bar{\mathbf{t}}_b$  and  $\mathbf{S}_{Tb}$  must be estimated under a model.

### 2.4 More than one model

The preceding development assumes that all estimands of interest are combinations of parameters of a single model. More generally, the contemplated analyses might involve fitting several models, and  $V$  might sum the variances of combinations of parameters from these models. An obvious

special case is estimation of a population mean (as suggested in Section 2.2), the coefficient of the model  $y_k = \beta_0 \cdot 1 + \varepsilon_{bk}$ , together with some regression coefficients. Another simple example is estimation of the means of variously defined domains, that is the coefficients of models of the form  $y_k = \beta'_{(m)} \mathbf{x}_{(m)bk} + \varepsilon_{(m)bk}$  where  $\mathbf{x}_{(m)bk}$  is a vector of domain membership indicators with alternative domain definitions indexed by  $m = 1, \dots, M$ , or contrasts of these means. For example, we might be interested in estimating mean outcomes both by race and by age.

If each of the combinations of interest only includes parameters of a single model, then each combination has its own design matrix, so the model index  $m$  can be identified with the estimand index  $i$ . Thus in (5) and (6) we replace  $\mathbf{X}_U$  with  $\mathbf{X}_{U(i)}$  and replace  $\mathbf{S}_b$  with  $\mathbf{S}_{(i)b}$ .

If some estimands combine parameters from different models, we stack the estimators  $\hat{\beta}_{(m)}$  for the different models. Then in (5) and (6) we replace  $\mathbf{X}'_U \mathbf{X}_U$  with  $\text{diag}(\mathbf{X}'_{U(m)} \mathbf{X}_{U(m)}, m = 1, \dots, M)$  and redefine  $\mathbf{S}_b$  as the combined sums of squares and crossproducts matrix for all of the models, with blocks

$$\mathbf{S}_{b(m', m'')} = \begin{pmatrix} \mathbf{u}_{b(m')} \mathbf{u}_{b(m'')}' & \mathbf{u}_{b(m')} \bar{\mathbf{t}}_{b(m'')} \\ \bar{\mathbf{t}}_{b(m'')} \mathbf{u}_{b(m')} & \mathbf{S}_{Tb(m', m'')} \end{pmatrix}.$$

The remainder of the optimization is unchanged from Section 2.3.

### 2.5 Flexible contrast weights

In CanCORS, we are interested in the income effect controlling for race and averaged across races. It is less important to us how the races are weighted in that average, since the study areas are not representative of national proportions by race. Then we might estimate the poverty/non-poverty income effect for each race and combine them with weights chosen to minimize the variance of the estimator of the weighted average of within-race income effects.

In general, we consider situations in which scientific interest is directed at estimating or testing *any* combination  $\mathbf{a}_i = \mathbf{A}_i \mathbf{f}_i$  where  $\mathbf{A}_i$  is fixed and each  $\mathbf{f}_i$  is arbitrary (and not necessarily all of the same dimension) subject to the constraints  $1' \mathbf{f}_i = 1$ ,  $f_{ij} \geq 0$ . In our motivating example, the underlying model includes eight indicator variables for each of the groups defined by four race groups crossed with dichotomous poverty level, and  $\mathbf{A}_i$  is an  $8 \times 4$  matrix in which each column contains a 1 and -1 for the contrast between poor and nonpoor within one race. Then  $\mathbf{f}_i$  contains the weights given to the contrast in each race, and  $\mathbf{a}'_1 = (f_{11}, -f_{11}, f_{12}, -f_{12}, f_{13}, -f_{13}, f_{14}, -f_{14})$  is the weighted contrast of the eight indicator coefficients.

Substituting into (5)-(6), we optimize over both sampling probabilities  $\pi = \{\pi_k\}$  and combining weights  $\mathbf{f} = \{\mathbf{f}_i\}$ . With multiple models, we use either of the formulations of Section 2.4, depending on whether the combinations of



interest include coefficients of one or several models. The definition of  $\mathbf{a}_i$  is thus determined in part by scientific considerations and in part by the information available from the population at hand.

A natural approach to jointly optimizing  $\pi$  and  $\mathbf{f}$  is alternately to minimize  $V$  with respect to  $\pi$  using the modified (6) and with respect to  $\mathbf{f}$ , observing the constraints on  $\mathbf{f}$ . In the optimization,  $\mathbf{f}_i$  appears in an expression of the form  $\mathbf{f}_i' \mathbf{D}_i(\pi) \mathbf{f}_i$ . Minimizing subject to the constraint  $\mathbf{f}_i' \cdot \mathbf{1} = 1$  using Lagrange multipliers, we obtain  $\hat{\mathbf{f}}_i = \mathbf{D}_i^{-1}(\pi) \mathbf{1} / (\mathbf{1}' \mathbf{D}_i^{-1}(\pi) \mathbf{1})$  as long as  $\pi_{hk} > 0$  and the nonnegativity constraints are not binding. If the nonnegativity constraints are binding, quadratic programming methods can be used.

## 2.6 Precision of unanticipated analyses

A design that is intended to be optimal for one regression coefficient might be very inefficient for other regression coefficients in the same or different models. Making the population mean one of the estimands helps to control this risk. We illustrate this by an example with design variables  $x_k, z_k$  with joint distribution

$$\zeta: (X, Z) \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$$

fully observed for individuals (indexed by  $k$  as in Section 2.2) and the following constant and univariate regression models:

$$\text{Model 0: } y_k = \alpha_0 + \varepsilon_k^0, \varepsilon_k^0 \sim [0, \sigma_0^2]$$

$$\text{Model 1: } y_k = \alpha_X + \beta_X x_k + \varepsilon_k^X, \varepsilon_k^X \sim [0, \sigma_X^2]$$

Model 2:  $y_k = \alpha_Z + \beta_Z z_k + \varepsilon_k^Z, \varepsilon_k^Z \sim [0, \sigma_Z^2]$ . To simplify notation we assume  $\sigma_0^2 \approx \sigma_X^2 \approx \sigma_Z^2 \approx 1$ ,  $\bar{x}_U = \bar{z}_U = 0$  and costs  $c_k$  are constant.

Consider the sample design optimized for  $V = dV(\hat{\alpha}_0) + V(\hat{\beta}_X)$ ,  $d \geq 0$ . By (4), the optimal design has  $\pi_k \propto \sqrt{d + x_k^2}$ . Under this design, the anticipated variance is approximated by  $V(\hat{\beta}_Z) \approx n^{-1} \sigma_Z^2 (Z_U' Z_U)^{-1} (Z_U' W_U Z_U) (Z_U' Z_U)^{-1}$  where  $Z_U = (z_1, \dots, z_N)'$  and  $W_U = \text{diag}(\pi_1^{-1}, \dots, \pi_N^{-1})$ . Then  $E_\zeta nV(\hat{\beta}_Z) \approx c_0 E_\zeta (Z^2 / \sqrt{d + X^2})$  where  $c_0$  depends only on  $d$  so  $E_\zeta nV(\hat{\beta}_Z)$  depends on  $\rho$  and  $d$ . If  $d = 0$  (no weight is attached to the estimation of mean),  $E_\zeta nV(\hat{\beta}_Z) = \infty$  unless  $\rho = \pm 1$ . Thus unless the objective gives some weight to the variance of the mean estimator, the design is potentially very poor for the coefficients attached to covariates that are not in the span of variables of the optimized models. But if  $d > 0$  we can decompose  $Z$  into components parallel and orthogonal to  $X$ ,  $Z = Z_1 + Z_2$  where  $Z_1 = \rho X$  and  $Z_2 = Z - \rho X$ , so  $Z_1 \perp Z_2$ ,  $Z_2 \perp X$  and  $E_\zeta Z_2 = 0$ . Then  $E_\zeta nV(\hat{\beta}_Z) = \rho^2 c_0 E_\zeta (X^2 / \sqrt{d + X^2}) + (1 - \rho^2) c_0 E_\zeta (1 / \sqrt{d + X^2}) = \rho^2 E_\zeta V_{\text{opt}}(\hat{\beta}_X) + (1 - \rho^2) E_\zeta V_{\text{opt}}(\hat{\alpha}_0)$ .

In words, the variance of the coefficient of the new model is a combination of the two variances that were controlled in the optimization. This suggests that a design that includes estimation of the overall population mean in the optimization gives some protection against extreme inefficiency for other models with variables that were not considered in the original design, although the simple results given here do not necessarily generalize to cover every case.

## 3. Application: Regressions on race and poverty status

### 3.1 Description of sites and data

The CanCORS project (Ayanian *et al.* 2004) consists of five geographically-defined sites (northern California, Los Angeles, Alabama/Georgia, North Carolina, and Iowa) and two organizationally-based sites. The northern California site consists of 9 counties extending from the San Francisco Bay area to semirural Placer County on the Nevada border. This site is ethnically diverse and geographically varied and therefore best illustrates the methods. We describe results for this site in detail and then summarize results for other sites.

Our data were based on the 2000 U. S. Decennial Census “long form” sample and were extracted for the 9 counties of our target area (Alameda, Contra Costa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara, and Solano) from SF-3, Tables 159a-159i, “Poverty Status in 1999 by Age.” We cross-tabulated the sampled residents at least 65 years old of each census block group (a small contiguous area roughly equivalent to several city blocks, henceforth referred to as a block) by race/ethnicity and income, using census sampling weights. The age restriction roughly corresponds to the ages of most incident cancer cases eligible for the study. Household income was dichotomized as exceeding or falling below the standard poverty line. The census included separate items on Hispanic ethnicity and race; we classified the population as Hispanic or as non-Hispanic white, Black, or Asian-American. A heterogeneous “Rest” category constitutes the remaining 3% of the elderly population. (For conciseness we henceforth refer to these as “race” categories.) The study site contained 844,560 over-65 individuals in 5,098 block groups, or an average of 166 per block group.

Table 1 summarizes the distribution of race and income in the northern California site. Blacks have the highest overall poverty rate and are also the most segregated (largest coefficient of variation of percent Black by block), consistent with national patterns of residential segregation (Denton and Massey 1993). Hispanics have the most relative geographical variation in poverty rates (largest coefficient of variation of poverty rates by block).

**Table 1**  
Distributions of race and poverty for those with age  $\geq 65$  years, by census block group in the northern California site. (CV = coefficient of variation)

	White	Black	Asian	Hispanic	Rest	Total
Percent of population	65.70	6.40	16.80	8.20	3.00	100.00
Percent poor	5.20	14.20	10.10	10.60	11.30	7.20
CV block percent in race	0.46	2.94	1.21	1.73	2.28	-
CV block percent poor	1.53	1.37	1.58	1.89	2.30	1.16

### 3.2 Design conditions: Available information and design objectives

We calculated the efficiency relative to simple random sampling (SRS) of the optimal design for scenarios defined by two conditions: (1) the choice of objective function, and (2) the assumptions about the information available for determining sampling probabilities.

We considered six possible assumptions about available information for race (unavailable, or available at the individual level) and income (unavailable, only available by block, or available at the individual level). Because race is more often recorded in hospital records than income, we excluded the case where individual income group is known but race is only known by block group. Each assumption corresponds to a definition of the cell for the development of Sections 2.3 and a corresponding definition of variables  $\mathbf{t}$  and  $\mathbf{u}$ :

1. No design information available: the cell is the entire population and  $\mathbf{u}$  includes race and income. (Columns headed "SRS" in Table 2.)
2. Race alone: the cell is a race category,  $\mathbf{u}$  contains race variables, and  $\mathbf{t}$  is income. (Columns headed "Race.")
3. Block-aggregated data alone: the cell is a census block group,  $\mathbf{u}$  is empty and  $\mathbf{t}$  includes race and income. (Columns headed "Block.")
4. Individual race, block-aggregated income data by race: the cell is the population of one race in a block group,  $\mathbf{u}$  is race, and  $\mathbf{t}$  is income. (Columns headed "Race+Block.")
5. Individual income, no race data: the cell is an income group,  $\mathbf{u}$  is income and  $\mathbf{t}$  is empty. (Columns headed "Income.")
6. Race and income both available for each individual: the cell is a race by income category,  $\mathbf{u}$  includes race and income, and  $\mathbf{t}$  is empty. (Columns headed "Race+Income.")

We calculated optimal sampling rates under each assumption about available information, with a variety of objective functions. Each of the objective functions we considered weights together variances of coefficient estimates in some or all of four regression models: (1) the "intercept only" model whose single parameter is the

population mean, (2) a race model parametrized as a white mean and contrasts for differences between whites and each of the other major race groups (Blacks, Hispanics, and Asians), (3) an income model parametrized as a nonpoor mean and a contrast between poor and nonpoor, and (4) an additive joint model including race and income effects. Every objective includes weight  $d_{\text{mean}} > 0$ , which guarantees that all  $\pi_{bk} > 0$ , avoiding numerical problems in the optimization. Thus, at least two models are represented in each objective (Section 2.4). When the objective weights both income and race effects, the single income effect is given weight  $d_{\text{income}} = 3$  to match the three race effects with weights of 1.

We explored a selection of objective weights that emphasized estimation of race effects, income effects, or both. Each panel of Table 2 represents a single choice of objective weights  $d_i$  (third column) for the contrast coefficients  $\mathbf{a}_i$  (second column) of a series of models (first columns). The fourth column shows the variance (normalized to unit sample size)  $nV_{\mathbf{a}_i}$  for estimation of that coefficient under SRS assuming residual variance  $\sigma^2 = 1$ . The remaining columns present design effects, the ratios of the normalized variance  $nV_{\mathbf{a}_i}$  for the optimized design with various assumptions about available design information to the variance under SRS. Rows with objective weight  $d_i = 0$  do not affect the optimization but are included to illustrate the effect of each design on efficiency for estimating a coefficient that is not included in the objective function. The final row summarizes the weighted design effect corresponding to the loss function, that is, the weighted combination of variances.

### 3.3 Efficiency with fixed models

The first two objective functions optimize for estimation of race contrasts and the overall mean. Using individual race greatly improves efficiency for estimating Black and Hispanic effects. The greatest gains are for the Black effect (the smallest of the three major racial minorities), whose variance is reduced to 43% of its value under SRS. Conversely there is no gain for Asian-Americans, whose population representation is close to the optimal sampling rate. With this objective, once race is available, additional design information (block or individual income) is irrelevant to the optimization. If individual race is unknown, using block of residence can help with oversampling of Blacks (the most segregated group residentially), reducing the variance of the estimated Black effect to about 65% of that under SRS, but oversampling by block only slightly reduces the variance of the estimated Hispanic effect. Knowing income by itself is of little use to improve sampling for estimation of race effects.

**Table 2**  
**Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions**

Objective 1: Optimized for race effects								
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.1	1.0	181	119	181	100	181
Race	Black	1	17.2	43	65	43	99	43
	Asian	1	7.5	100	106	100	100	100
	Hispanic	1	13.7	55	90	55	100	55
Income	Poor	0	15.0	182	104	182	81	182
Race+Income	Black	0	17.4	44	65	44	99	44
	Asian	0	7.5	100	106	100	100	100
	Hispanic	0	13.8	55	90	55	100	55
	Poor	0	15.2	182	104	182	81	182
Total = $nV = n\sum d_i V_{a_i}$			38.6	59	82	59	99	59

Objective 2: Optimized for race effects and overall mean								
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	136	115	136	100	136
Race	Black	1	17.2	44	66	44	99	44
	Asian	1	7.5	100	104	100	100	100
	Hispanic	1	13.7	56	90	56	100	56
Income	Poor	0	15.0	121	101	121	82	121
Race+Income	Black	0	17.4	45	66	45	99	45
	Asian	0	7.5	100	104	100	100	100
	Hispanic	0	13.8	56	90	56	100	56
	Poor	0	15.2	122	102	122	82	122
Total = $nV = n\sum d_i V_{a_i}$			41.5	65	84	65	100	65

Objective 3: Optimized for income effect								
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	0.001	1.0	103	154	173	173	173
Race	Black	0	17.2	75	119	152	163	163
	Asian	0	7.5	90	144	173	170	170
	Hispanic	0	13.7	86	142	196	168	168
Income	Poor	3	15.0	97	74	60	27	27
Race+Income	Black	0	17.4	75	119	153	164	164
	Asian	0	7.5	90	144	174	171	171
	Hispanic	0	13.8	86	143	197	169	169
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			45.0	97	74	60	27	27

Objective 4: Optimized for income effect and overall mean								
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Variance as percent of variance under SRS (by available design information)				
				Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	134	147	151	151
Race	Black	0	17.2	76	107	128	142	142
	Asian	0	7.5	91	127	145	148	148
	Hispanic	0	13.7	86	125	161	147	147
Income	Poor	3	15.0	97	75	61	27	27
Race+Income	Black	0	17.4	76	107	129	143	143
	Asian	0	7.5	91	127	146	149	149
	Hispanic	0	13.8	86	125	162	147	147
	Poor	0	15.2	97	75	63	29	29
Total = $nV = n\sum d_i V_{a_i}$			48.0	97	79	66	35	35



Table 2 (continued)

Normalized variances and objective functions for optimal designs for various objective weights and design information assumptions

Objective 5: Optimized for separate race effects, income effect and overall mean								
			Variance as percent of variance under SRS (by available design information)					
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	117	135	114	150
Race	Black	1	17.2	54	74	55	109	53
	Asian	1	7.5	95	106	109	112	116
	Hispanic	1	13.7	67	96	69	112	67
Income	Poor	3	15.0	101	82	72	38	37
Race+Income	Black	0	17.4	55	74	55	109	52
	Asian	0	7.5	95	106	109	113	115
	Hispanic	0	13.8	67	96	69	112	66
	Poor	0	15.2	101	82	72	39	35
Total = $nV = n\sum d_i V_{a_i}$		0	86.4	86	86	74	73	56

Objective 6: Optimized for race effects and income effect in two-factor model and for overall mean								
			Variance as percent of variance under SRS (by available design information)					
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	111	119	138	114	156
Race	Black	0	17.2	55	74	55	108	53
	Asian	0	7.5	95	107	109	112	114
	Hispanic	0	13.7	67	96	69	111	67
Income	Poor	0	15.0	101	82	72	38	37
Race+Income	Black	1	17.4	55	74	55	109	52
	Asian	1	7.5	95	107	109	113	113
	Hispanic	1	13.8	67	96	69	112	66
	Poor	3	15.2	101	81	71	39	35
Total = $nV = n\sum d_i V_{a_i}$			87.2	86	86	73	73	54

Objective 7: Optimized for income effect in two-factor model and for overall mean								
			Variance as percent of variance under SRS (by available design information)					
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	103	135	149	148	156
Race	Black	0	17.2	77	100	98	139	97
	Asian	0	7.5	91	124	132	145	132
	Hispanic	0	13.7	86	122	135	144	122
Income	Poor	0	15.0	97	75	62	28	28
Race+Income	Black	0	17.4	77	100	98	140	96
	Asian	0	7.5	91	124	132	146	132
	Hispanic	0	13.8	86	122	135	144	121
	Poor	3	15.2	97	75	62	29	27
Total = $nV = n\sum d_i V_{a_i}$			48.6	97	79	67	36	35

Disproportionate sampling, tuned to optimize for estimation of race effects, inflates the variances of the other parameter estimators. When minimal weight is given to the mean in the optimization objective (Objective 1), this inflation can be quite large: a factor of 181% for the mean and income effects. Giving more weight to the mean (Objective 2) moderates this effect, reducing the variance inflation to 136% for the mean and 121% for the income effect, while only slightly increasing variances for the race effects.

The minimum possible normalized variance for estimation of the income effect (Objective 3) is 4 (27% of the variance under SRS), attained when income is known for individuals under a design that divides the sample equally between poor and nonpoor. With block-level information, variance can be reduced to 74% of that under SRS. Although knowing race alone has little benefit for this objective, adding individual race to block-level information further reduces the variance of the estimated income effect to 60% of that under SRS. Variances of estimates of the mean and of race effects are substantially increased under

these designs, but increasing the weight of the mean (Objective 4) substantially ameliorates the variance inflation for the mean and race effects, only slightly increasing the variance of the estimated income effect.

Including both race and income effects in Objective 5 yields designs that are not quite as good as the optimal designs for either alone, but still much better than SRS. For example, variances of the race effects with race and block of residence known are 10% to 24% higher than with the designs using the same design information but separately optimized for race or income. When only individual race or only individual income is known, the design essentially optimizes for the effects corresponding to the available variable, inflating the variance of estimated effects of the other variable.

The design optimized for joint race and income effects in the two-factor additive model (Objective 6) is quite close to that optimizing for race and income effects in separate marginal models (Objective 5). When optimizing for separate effects, variances of these effects are slightly smaller than those of the corresponding effects in the two-factor model. When optimizing for effects in the joint model, their variances are reduced although in most cases still slightly larger than those of the corresponding effects in marginal race and income models, due to the partial confounding of race and income effects.

Likewise, optimization for the income effect in the two-factor model (Objective 7) is fairly similar to optimization for the univariate income effect (Objective 4) when no race data are available. Making race data available together with either block or individual income, however, considerably reduces variances for race effects under the design for the two-factor model. Because of the partial confounding of race and income effects under this model, this design adapts to estimate the former more efficiently, accumulating more data at the design points that are critical to unconfounding these effects.

3.4 Efficiency with flexible contrast weights

We next consider the potential benefits of estimating income effects under a flexible weighting scheme (Table 3). The objective function considers coefficients of two models, the constant model whose parameter is the population mean, and a model with indicator variables for each race-by-income cell. The income effect within each race is estimated as the difference of the coefficients for poor and nonpoor within that race, and these estimates are combined with flexible weights to estimate an overall income effect. This strategy is most nearly parallel to Objective 7, which also estimates income effects controlling for race. The flexible-contrast analysis is less model-dependent than the two-factor model in that it does not rely on that model's

additivity assumption. On the other hand, the way the races are combined does not necessarily reflect population proportions. The weights given to the income contrast in each race, estimated as described in Section 2.5, are presented in the lower panel of Table 3 to demonstrate how this approach allows us to modify the estimand to exploit available design information. (The alternating-optimization algorithm converged to adequate accuracy within 7 iterations.)

Under SRS the variance of the income effect under the flexible-weights model is slightly larger than in the two-factor model (15.91 versus 14.99). The weight given to the white contrast under this design (51%) is less than the white proportion of the population (66%) because relatively few whites are poor and therefore the income contrast among whites is relatively imprecise. Conversely, the weight for the Black income contrast (12%) is almost twice that group's share of the population, because of the disproportionately high poverty rates in that group.

Using individual race in the design accentuates this disproportion: more sample, and much more weight (75%), is given to the Black group, with the highest percentage in poverty. Thus flexible weighting makes possible a large reduction in the variance of the estimated income effect (to 63% of that under SRS) using only race, which was not possible under the more restrictive two-factor DPQ model.

Block-level information is slightly less useful for this design than race information. The combination of block and race information, however, is very powerful, reducing the variance of the income effect to 48% of that under SRS. Under this design, much more weight (46%) is given to the Hispanic income contrast, which can be estimated efficiently because of the greater income segregation among Hispanics (Table 1). When individual income is available (with or without race), the contrasts weights approximate the proportions by race, since efficient income contrasts can be obtained within any race and the inclusion of the overall mean in the objective pulls the design toward proportionate sampling. Thus, the design is dramatically different under alternative assumptions about availability of design information.

3.5 Comparisons across sites

Table 4 compares the gains for disproportionate sampling at four CanCORS sites, excluding the nongeographical sites and one site (Iowa) that was almost all white. At each site we optimized for unit ( $d_i = 1$ ) weighting of variances of overall mean and the income effect in the two-way model (proportional to Objective 7), under alternative assumptions about available design information. The theoretical minimum for this objective with a balanced population is 5 ( $V_{\text{mean}} = 1, V_{\text{income}} = 4$ ). SRS is inefficient at every site,

especially in Alabama and northern California, and race information alone would be of little help. Conversely, the best variance attainable using full race and income information on individuals is between 5.60 and 5.72 at each site. Oversampling based on block-level income information would substantially reduce variances, with substantially greater gains in Alabama and northern California than in the other sites.

4. Discussion

To develop design alternatives for a health services study, we extended previous methods for optimal design in domain estimation to show how an optimal unequal-probability sampling scheme can be designed for estimation

of regression coefficients in one or more models. In our application, substantial reductions in variance were possible even if some variables were only available for geographical aggregates. Particularly large gains were possible for categorical regressors (poverty status, race) with very imbalanced distributions.

In essence, our approach to survey design with imprecisely measured design variables uses the predictive distribution of the design variables for each sampled unit, specifically the expectations of the variables and of their squares and cross-products. This concept unites design using cell aggregates (estimated from census or sample data), using variables measured with error, or using a sampling frame whose units might have changed their characteristics over time.

**Table 3**  
Normalized variances and contrast weights for optimal DPQ designs with flexible weighting of income contrasts by race. Lines for fixed contrasts are included to demonstrate the effect of various choices of flexible weights, for comparison to fixed-weight objective scenarios

Variances for “flexible-weight” estimate and for contrasts represented in Table 2								
Variance as percent of variance under SRS (by available design information)								
Model	Effect	Weight ( $d_i$ )	Variance Under SRS	Race	Block	Race+ Block	Income	Race+ Income
Constant	Mean	3	1.0	233	139	206	152	152
Flexible contrast	Income	3	15.9	63	74	48	28	26
Race	Black	0	17.2	34	90	61	143	139
	Asian	0	7.5	197	124	172	149	146
	Hispanic	0	13.7	172	124	61	148	144
Income	Poor	0	15.0	209	77	124	27	39
Race+Income	Black	0	17.4	35	90	61	144	139
	Asian	0	7.5	197	124	173	150	147
	Hispanic	0	13.8	172	124	61	148	144
	Poor	0	15.2	210	77	125	29	39
Total = $\sum d_i V_{a_i}$			50.7	73	78	57	35	33

Optimum weights (as percent,  $100\% \times f_i$ ) of each within-race income contrast in calculation of the combined estimate of the income effect, under each design information assumption. (Columns may not sum to 100% due to roundoff error.)

Contrast	Design information assumptions					
	SRS	Race	Block	Race+Block	Income	Race+Income
Black	12	75	17	25	9	6
Asian	24	9	26	16	21	17
Hispanic	12	5	13	46	11	8
White	51	12	45	13	59	68

**Table 4**  
Normalized objective function for optimal DPQ designs under equal ( $d_i = 1$ ) weighting of variances of the overall mean and the income effect in the two-way model, at four CanCORS sites

Site location	Variance under SRS	Variance as percent of variance under SRS				
		Race	Block	Race+Block	Income	Race+Income
Alabama	16.2	97	79	67	36	35
Los Angeles	11.8	98	85	76	49	47
North Carolina	10.2	97	89	86	59	55
Northern California	16.2	97	79	67	36	35



The methods described here for optimizing element sampling probabilities can be combined with stratification and cluster or multistage sampling. (Neither of these design features appeared in the CanCORS study which motivated our research. Stratification was inconvenient given the sequential identification of subjects and there was little prior information to guide construction of homogeneous strata. Telephone interviewing made it operationally unnecessary to cluster our subjects.) Because these design features can affect the sampling distributions of both the design and outcome variables, and the design objectives involve both the posited population model and the scientific model of interest, the number of possible combinations is even larger than in design for estimation of a population mean. We therefore limit ourselves to suggesting a few ideas to be followed up in future research.

Stratification can improve a design for a regression analysis in at least three ways: (1) to implement disproportionate sampling (using probabilities equal or close to those derived under our methodology), (2) to control the distribution of design variables to be closer to the optimal design than in an unstratified unequal-probability design, and (3) to reduce the within-stratum variation of the case influence statistics and thereby reduce the variance of coefficient estimates (Fuller 1975). Since the efficiency of the design is insensitive to small deviations around the optimum, some stratified designs with equal probabilities within strata might approach the efficiency of the optimal design. *Ad hoc* stratifications might have poorer efficiency, even with optimal allocation to strata. For example, stratifying blocks by the least prevalent race-income group represented yielded a design with about half the efficiency gain of our design using aggregated block composition.

With regard to the last point, note that designing homogeneous strata for estimation of regression coefficients is likely to be more difficult than for estimation of a mean. The influence of an observation depends on its residual from the regression model, not its raw value, so to reduce homogeneity the stratification would have to involve predictive variables not included in the model. Influence also depends on the observation's leverage for each coefficient, a possibly complex function of the covariates.

For cluster sampling, the equivalence of  $E_{\pi} \text{Var}_{\pi} \mathbf{a} \hat{\beta}$  and  $E_{\pi} \text{Var}_{\pi} \mathbf{a} \hat{\beta}$  might not hold except under restrictive assumptions such as independent residuals; thus the terms of the middle factor of (5) would take a more complex form. There are several possible cases for cluster sampling depending on the relationship between the cells and the clusters, which should be elaborated on further research.

Another natural extension is to nonlinear regression models and other estimands defined by estimating equations. The weighted least squares formulation of the

Newton-Raphson step (McCullagh and Nelder 1989, sec. 2.5) for a generalized linear model can be applied by suitably defining  $\sigma_{\beta k}^2$  in (3) and hence in (4)-(6); a similar procedure can be applied for other estimating equations (Binder 1981; Binder 1983; Morel 1989). Because the variances are functions of the model predictions, implementing this modification requires design assumptions about the fitted model as well as about the distribution of the covariates.

Every optimization has its costs, which for our methods can be both practical and statistical.

In the CanCORS study, incident cases of the cancers under study were identified in real time through a field operation ("rapid case ascertainment"); patients then had to be contacted on a very tight schedule to start contacting them for interviews within the desired interval (3 months from their dates of diagnosis). Thus, the practical issues of survey implementation were exacerbated. Among the concerns that ultimately led us not to implement the DPQ design were (1) the difficulty of accurately geocoding patients within the time frame allowed; (2) incomplete and inaccurate race identification in the case ascertainment data, and (3) lower-than-expected participation rates, which made any sampling problematical.

Such issues are less problematic in surveys with a static sampling frame that can be processed on a less stringent timeline, particularly in large-scale and/or repeated surveys in which even modest variance reductions justify some added complexity. They could be used, for example, to evaluate the potential gains through geographically-based oversampling in surveys for which national estimates by race are required.

Statistical concerns about our design strategy arise because optimization for one set of predetermined statistical objectives is likely to reduce efficiency for others. It is difficult in any but the most tightly focused study to anticipate all potential analyses. Simultaneous optimization for a reasonably comprehensive collection of analyses, and investigation of sensitivity of the design to varying the relative weights of the various objectives, should give some protection against an overspecialized design. However, this approach can only be used with variables for which there are some data prior to the study. The results in Section 2.6 suggest that monitoring the effect of disproportionate sampling on the precision of the population mean gives some protection against designs that are excessively inefficient for unanticipated analyses and variables, although the bounds there are not very general.

More broadly, we might ask when the DPQ analysis is the scientifically relevant estimand. Regression models are often used in analyses intended to be generalizable to broader populations, rather than to describe the finite

population at hand, just as the CanCORS sites were selected purposively to study patterns and variations in care that might reflect broader national patterns. While using sampling weights in enumerative studies is relatively uncontroversial, there has been a lively debate about the use of weights in analytic studies (Hansen, Madow and Tepping 1983 and discussion; DuMouchel and Duncan 1983; Bellhouse 1984; Pfeffermann 1993, Fuller 2002, sec. 5). A population-descriptive analysis offers some robustness against the possibility that the sample will be selected in way that distorts typical relationships. Thus, even where a pure DPQ analysis cannot be justified on grounds of enumerative representativeness, a sample drawn to optimize unweighted estimation of regression coefficients might have limited scientific value. For example, suppose that the CanCORS data would be analyzed with an *unweighted* regression to estimate a simple income effect (a contrast of means), using block level design information from the census. Optimally the sample would draw from a collection of blocks which, taken together, have about half their residents in poverty. Since poverty rates are rarely that high, this effectively requires sampling only from the blocks with the highest poverty rates. Such a sample would be unrepresentative of either of the income groups. Similarly, a sample that overrepresented Black residents by sampling from mostly Black blocks would (if analyzed without weights) be unrepresentative of the Black population in general, because the services available in highly segregated areas are likely to differ from those in more mixed areas.

More general formulations are needed, with clearly stated assumptions and objectives, that “consider[s] the model parameters as the ultimate target parameters but at the same time focuses on the DPQ’s as a way to secure the robustness of the inference” (Pfeffermann 1993), taking into account the scientific objectives of the study. Previous proposals include testing the null hypothesis that the weights have no effect on the regression (DuMouchel and Duncan 1983; Fuller 1984), or including design variables (Nathan and Holt 1980; Little 1991) or the weights themselves (Rubin 1985) as control variables in the regression. These approaches are problematical, however, when the weights are functions of the covariates of primary scientific interest. We have attempted through flexible contrast weighting (Section 2.5) to take a step toward such a general formulation, extending the DPQ approach to allow a focus on a range of valid inferences for particular scientific objectives rather than exclusively on inference for finite populations. From this range, the investigator can select an inferential objective and sample design adapted to the structure of the population and the practicalities of study design.

## Appendix

### Equivalence of sampling and model variances

We show that  $E_{\xi} \text{Var}_{\pi} \hat{\beta} \approx E_{\pi} \text{Var}_{\xi} \hat{\beta}$  under the following conditions:

1.  $1/N \mathbf{X}'_U \mathbf{X}_U \rightarrow \Sigma$  for some positive definite  $\Sigma$ . This minimal condition relates the hypothetical sequence of populations.
2. The design-based regression estimator can be written as  $\hat{\beta} = \beta_U + R_{n,S}$  where  $\text{Var}_{\xi} E_{\pi} R_{n,S} = o(n^{-1})$  and  $\text{Var}_{\pi} E_{\xi} R_{n,S} = o(n^{-1})$ . Note that  $\hat{\beta}$  cannot strictly be defined as in (2), because the matrix inverse is undefined when the sample values of  $x$  do not span the design space and hence its expectation and variance are also undefined. A scalar ratio estimator likewise might be undefined with nonzero but  $o(n^{-1})$  probability because the sample might have only 0 values for the denominator variable. Assigning some arbitrary value in that event, the estimator nonetheless could have good asymptotic properties. A similar argument lets us assume that a suitable  $\hat{\beta}$  can be defined. We do not specify how (2) must be modified to technically satisfy the condition since this depends on the specifics of  $\xi$  and the sequence of designs.
3.  $\max(\pi_i) = O(n/N)$  and  $n = o(N)$ , essentially our assumption that finite population corrections can be ignored.
4. Homoscedasticity,  $\text{Var}_{\xi} y_k = \sigma^2$ ; this is not restrictive since it can always be made true by a suitable transformation of  $x$  and  $y$ .

Under these conditions,

$$\begin{aligned} \text{Var}_{\pi\xi}(\hat{\beta}) &= \text{Var}_{\pi} E_{\xi} \hat{\beta} + E_{\pi} \text{Var}_{\xi} \hat{\beta} \\ &= \text{Var}_{\pi} \beta + E_{\pi} \text{Var}_{\xi} \hat{\beta} \\ &= o(n^{-1}) + E_{\pi} \text{Var}_{\xi} \hat{\beta} \end{aligned}$$

On the other hand

$$\text{Var}_{\xi\pi} \hat{\beta} = \text{Var}_{\xi} E_{\pi} \hat{\beta} + E_{\xi} \text{Var}_{\pi} \hat{\beta}$$

The first term in the above equation is

$$\begin{aligned} \text{Var}_{\xi} E_{\pi}(\beta_U + R_{n,S}) &= \text{Var}_{\xi}(\beta_U + E_{\pi} R_{n,S}) \\ &= (\mathbf{X}'_U \mathbf{X}_U)^{-1} \sigma^2 + o(n^{-1}) + o(N^{-1/2} n^{-1/2}) \\ &= O(N^{-1}) + o(n^{-1}) + o(n^{-1/2} N^{-1/2}) = o(n^{-1}) \end{aligned}$$

This proof is an elaboration of one by Isaki and Fuller (1982), summarized in Pfeffermann (1993, page 321).



## Acknowledgements

This research was funded by grants U01-CA93344 (Zaslavsky and Zheng), U01-CA93324 (Zaslavsky), and U01-CA093348 (Adams) from the National Cancer Institute. The authors thank Nat Schenker and Van Parsons for useful comments on an earlier draft, and the associate editor and two referees for thoughtful comments.

## References

- Ayanian, J.Z., Chrischilles, E.A., Wallace, R.B., Fletcher, R.H., Fouad, M., Kiefe, C.I., Harrington, D.P., Weeks, J.C., Kahn, K.L., Malin, J.L., Lipscomb, J., Potosky, A.L., Provenzale, D.T., Sandler, R.S., Vanryn, M. and West, D.W. (2004). Understanding cancer treatment and outcomes: The Cancer Care Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 2, 2992-2996.
- Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.
- Binder, D.A. (1981). On the variances of asymptotically normal estimators from complex surveys. *Survey Methodology*, 7, 157-170.
- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Botman, S.L., Moore, T.F., Moriarity, C.N. and Parsons, V.L. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004*. Vital and Health Statistics, 2(130). Washington, DC: National Center for Health Statistics.
- Cassel, C.M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- Chromy, J.R. (1987). Design optimization with multiple objectives. In *Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA. 194-199.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.
- Denton, N.A., and Massey, D.S. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA: Harvard University Press.
- Dumouchel, W.H., and Duncan, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.
- Eaton, D.K., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Harris, W.A., Lowry, R., Mcmanus, T., Chyen, D., Shanklin, S., Lim, C., Grunbaum, J.A. and Wechsler, H. (2006). Youth risk behavior surveillance - United States, 2005. *Morbidity and Mortality Weekly Report*, 55(SS-5), 1-108.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C37, 117-132.
- Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, 10, 97-118.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.
- Hansen, M.H., Madow, W.G. and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.
- Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-97.
- Kish, L. (1974). Optimal and proximal multipurpose allocation. Alexandria, VA. In *Proceedings of the Social Statistics Section*, American Statistical Association, 111-118.
- Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (Second Edition). London: Chapman & Hall Ltd.
- Morel, J.G. (1989). Logistic regression under complex survey designs. *Survey Methodology*, 15, 203-223.
- Morris, C., Newhouse, J.P. and Archibald, R. (1979). On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments. *Research in Experimental Economics*, 1, 199-220.
- Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B, Methodological*, 42, 377-386.
- Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith). Amsterdam: Elsevier/North-Holland, 463-472.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.
- Sastry, N., Ghosh-Dastidar, B., Adams, J. and Pebley, A.R. (2005). The design of a multilevel survey of children, families, and communities: The Los Angeles Family and Neighborhood Survey. *Social Science Research*, 35, 1000-1024.
- Thomsen, I., Tesfu, D. and Binder, D.A. (1986). Estimation of design effects and intraclass correlations when using outdated measures of size. *International Statistical Review*, 54, 343-349.
- Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*, New York: John Wiley & Sons, Inc.



- Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. In *ASA Proceedings of the Social Statistics Section*, American Statistical Association. Alexandria, VA.
- Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. In *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, 497-502.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.

# Generalized regression estimators of a finite population total using the Box-Cox technique

Yan Li<sup>1</sup>

## Abstract

A new generalized regression estimator of a finite population total based on the Box-Cox transformation technique and its variance estimator are proposed under a general unequal probability sampling design. By being design consistent, the proposed estimator maintains the robustness property of the GREG estimator even if the underlying model fails. Furthermore, the Box-Cox technique automatically finds a reasonable transformation for the dependent variable using the data. The robustness and efficiency of the new estimator are evaluated analytically and via Monte Carlo simulation studies.

**Key Words:** Generalized regression (GREG) estimator; Box-Cox technique; Pseudo-maximum likelihood (PML).

## 1. Introduction

Generalized regression (GREG) estimators for finite population totals and means are derived using suitable regression models. Although models are employed to construct such estimators, randomization must be used to select the samples and to evaluate the statistical properties of the resulting estimation strategies. Examples may be found in Särndal, Swensson and Wretman (1992), Estevao, Hidiroglou and Särndal (1995), Fuller, Loughin and Baker (1994), and Jayasuriya and Valliant (1996). A good model is crucial in limiting the variability of a *model-assisted* estimator like the GREG. If the assumed model describes the finite population well, the GREG estimator can potentially bring about a large variance reduction when used in place of the Horvitz-Thompson estimator (Horvitz and Thompson 1952). A general discussion of regression estimation can be found in Fuller (2002). Särndal *et al.* (1992) provide a comprehensive description of the model-assisted framework for constructing survey estimators.

Studies on the GREG estimator have mostly been conducted in the context of linear regression modeling. The GREG essentially incorporates relevant auxiliary variables through their known population control totals even when the auxiliary variables are known for every unit in the population (Cassel, Särndal and Wretman 1976; Särndal 1980; Deville and Särndal 1992; Särndal *et al.* 1992; Jiang and Lahiri 2006). The availability of complete auxiliary information is fairly common these days: census data, administrative registers, remote sensing data and previous surveys provide a wealth of valuable information that can be used to increase the precision of the estimation procedure (Montanari and Ranalli 2003). As a result, complex models and flexible techniques making use of complete auxiliary information have been introduced into survey sampling in

recent years. Penalized spline techniques have been adapted to construct model-based (Zheng and Little 2004) and model-assisted (Breidt, Claeskens and Opsomer 2005) estimators for a finite population total based on complex survey data. Breidt and Opsomer (2000) considered a nonparametric, model-assisted regression estimator using local polynomial regression and showed that nonparametric regression can significantly improve the efficiency of estimators when parametric models are misspecified. Their work was further extended from the single-covariate model to the case of the semiparametric additive model. Wu and Sitter (2001) fit a general working model, which could have both linear and nonlinear components, and then calibrated on the resulting fitted values using simple linear regression. Montanari and Ranalli (2005) combined model calibration estimation and nonparametric methods and proposed nonparametric model-assisted estimators for a finite population mean.

In mainstream statistics, a suitable transformation on the dependent variable in the assumed model is often taken to achieve normality, linearity, and homoscedasticity (Carroll and Ruppert 1988), but the literature on transformations in finite population inference is not very rich. There is, however, a growing interest in developing methods that use an appropriate transformation with survey data. In some survey applications, especially in business and establishment surveys, it is common to have highly skewed continuous and positive survey variables (*e.g.*, income). To estimate the finite population total of the survey variable, a linear model may not be appropriate for a study variable, but may be reasonable for a strictly monotonic transformation of the study variable. Chen and Chen (1996) considered transformed survey data in order to improve on the precision of the normal approximation. Korn and Graubard (1998) compared different confidence intervals, including intervals

1. Yan Li, University of Maryland at College Park. Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS; Executive Plaza South; Room 8014, 6120 Executive Boulevard, MSC7244, Bethesda, MD 20892, U.S.A. E-mail: lisherry@mail.nih.gov.

based on a logit-transformation, for proportions with small expected number of positive counts. Karlberg (2000) proposed an estimator based on a lognormal-logistic super-population model to predict the finite population total of a highly skewed survey variable. The simulation results indicated that the lognormal-logistic model estimator offers a sensible alternative to other estimators, especially when the sample size is small. Chambers and Dorfman (2003) discussed the estimation of a finite population mean under certain general but known transformation on the continuous data.

Researchers find the transformation technique useful in analyzing survey data. The key step is the identification of an appropriate transformation that fits the survey data well. In many applications, the form of transformation is determined subjectively. Unfortunately, prior knowledge or theory may not suggest the transformation to be used. In such situations, it would be convenient to determine the transformation adaptively using the survey data.

The work of Box and Cox (1964) has led to the development of "data-decide-transformation" methods for constructing models with independently and identically distributed errors. Their paper and other papers on the subject, including Tukey (1957), John and Draper (1980), and Bickel and Doksum (1981), have inspired a large volume of applied research. Spitzer (1976) examined the relationship between the demand for money and the liquidity trap with a generalized Box-Cox model. In the context of research related to malaria, Newman (1977) concluded that the Box-Cox functional specification was superior to earlier specifications. Miner (1982) and Davison, Arnade and Hallahan (1989) considered modeling of soybean yield functions and the U.S. soybean export respectively. They concluded that the Box-Cox transformation provides approximately normally distributed error terms. A bibliography related to the Box-Cox transformation can be found in a review paper by Sakia (1992). For an application of the Box-Cox methodology to a mixed linear model, see Gurka (2004, 2006).

Li and Lahiri (2007) used the Box-Cox transformation on the study variable to generate robust model-based predictors of a finite population total. Model-assisted estimators were also mentioned in a sub-section (Section 2.6), but the properties of the proposed estimators were not investigated. This article provides that analysis. The Box-Cox technique is employed to fit a regression between the study variable and a set of auxiliary variables. The fitted regression is then used to predict the values of study variable for the unobserved units of the finite population which, in turn, provide an adaptive regression type estimator within the model-assisted framework.

The article is organized as follows. In Section 2, a new estimator is proposed and the analytical properties of the proposed estimator with respect to the sampling design are investigated. To better assess the robustness and the efficiency of the proposed estimator, we compare it to GREG estimators based on the underlying linear and log-linear working models via Monte Carlo simulations in Section 3. Section 4 discusses the results. Finally, we offer some concluding remarks about areas for potential future research in Section 5.

## 2. A new estimator of the finite population total

Suppose that the quantity of interest is the finite population total

$$T = \sum_{i \in U} y_i,$$

where  $U = \{1, \dots, N\}$  denotes a finite population of known size  $N$ , and  $y_i > 0$  is the value of the study variable associated with unit  $i$ . Write  $\mathbf{Y} = (y_1, \dots, y_N)'$ . To estimate  $T$ , a sample  $s$  of size  $n$  is drawn from the finite population using a probability sampling scheme. Let  $w_i$  be the sampling weight for unit  $i$ . The sampling weight is simply the inverse of the inclusion probability for the unit  $i$ , denoted by  $\pi_i = P(i \in s)$  ( $i = 1, \dots, N$ ). We assume that we have information on  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$ , where  $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ik})'$  is a column vector of  $k$  known auxiliary variables for the unit  $i$ . For any sample  $s$  of size  $n$ , we redefine  $\mathbf{Y}$  and  $\mathbf{X}$  so that the first  $n$  rows of  $\mathbf{Y}$  and  $\mathbf{X}$  correspond to those in the sample. Write

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_s \\ \mathbf{y}_r \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{pmatrix},$$

where

- $\mathbf{y}_s$  is a  $n \times 1$  column vector of observed study variable;
- $\mathbf{y}_r$  is a  $(N - n) \times 1$  column vector of unobserved study variable;
- $\mathbf{X}_s$  is a  $n \times (k + 1)$  matrix of known auxiliary variables in the sample;
- $\mathbf{X}_r$  is a  $(N - n) \times (k + 1)$  matrix of known auxiliary variables outside the sample.

Throughout the paper, we use  $E_d$  and  $V_d$  to denote the expected value and variance with respect to the sampling design.

### 2.1 GREG estimators of finite population totals

The GREG estimator is defined here as

$$\hat{T}_G = \sum_{i \in U} \hat{y}_{i,w} + \sum_{i \in s} (y_i - \hat{y}_{i,w}) / \pi_i,$$



where  $\hat{y}_{i,w}$  is the predictor of  $y_i$  based on a model. Regardless of how well the underlying model describes the population, GREG estimators of the finite population total are design-consistent under mild conditions (Särndal *et al.* 1992). The most commonly used model is the standard linear regression model, given by

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where  $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$ , a  $N$ -variate probability distribution with the mean vector  $\mathbf{0}$  and variance covariance matrix  $\sigma^2 \mathbf{I}$ , and  $\mathbf{I}$  is the  $N \times N$  identity matrix (nothing would be lost in this context by replacing  $\sigma^2 \mathbf{I}$  with a more general positive definite matrix). In this equation,  $\boldsymbol{\beta}$  is a  $(k+1) \times 1$  column vector of regression coefficients. Both  $\sigma^2$  and  $\boldsymbol{\beta}$  are unknown superpopulation parameters. An unbiased predictor for the  $i^{\text{th}}$  unit is

$$\hat{y}_{i,w} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w, \quad (1)$$

where  $\hat{\boldsymbol{\beta}}_w$  is the weighted least square (WLS) estimator of  $\boldsymbol{\beta}$  under  $\mathbf{M}_1$  and

$$\hat{\boldsymbol{\beta}}_w = \left( \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i \in s} w_i \mathbf{x}_i y_i \right).$$

In some applications, especially in business and agricultural surveys, a linear model may not be appropriate for  $y$ , but may be reasonable for a strictly monotonic transformation of  $y$ . For the data set given in Royall and Cumberland (1981), Chen and Chen (1996) observed that the finite population distribution was severely skewed and that the log-transformation helped achieving symmetry. The need and the benefit of taking the log-transformation were obvious. Therefore, we consider the log-linear regression model where the log-transformation is used on the dependent variable

$$\mathbf{M}_2 : \log \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where  $\boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I})$ . An obvious predictor for the  $i^{\text{th}}$  unit is given by

$$\hat{y}_{i,w} = e^{\mathbf{x}_i' \hat{\boldsymbol{\beta}}_w}, \quad (2)$$

where

$$\hat{\boldsymbol{\beta}}_w = \left( \sum_{i \in s} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i \in s} w_i \mathbf{x}_i \log y_i \right).$$

Model  $\mathbf{M}_2$  requires a subjective specification of the transformation applied to the study variable. This may be reasonable in situations where we know the appropriate transformation from prior empirical evidence or from the theory. In absence of any prior knowledge about the transformation, it is prudent to choose the transformation

from among a flexible family of transformations using the data.

Tukey (1957) considered the following family of power transformations:

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

where  $y > 0$ . In order to remove the discontinuity at  $\lambda = 0$ , Box and Cox (1964) proposed the following family of transformations:

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

where  $y > 0$ . The parameter  $\lambda$  determines the nature of transformation. For example,  $\lambda = 1, 0, 0.5, -1$  correspond to no transformation, log-transformation, square root transformation, and reciprocal transformation, respectively. The transformation parameter  $\lambda$  is estimated by the data. The Box-Cox analysis may lead to a log-transformation, but may equally lead to some other transformation in the above family - it depends on the actual data observed.

We consider the following superpopulation model for the transformed study variable:

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where  $\boldsymbol{\varepsilon}$  are approximately normal with mean  $\mathbf{0}$  and variance matrix  $\sigma^2 \mathbf{I}$ .

Schlesselman (1971) showed that the maximum likelihood estimators of the Box-Cox model parameters are scale-invariant so that rescaling the original  $y$ -variable leads to the same log-likelihood function as long as the regression model contains an intercept term. Following his study and most papers on the Box-Cox models, we include an intercept term in the model.

Under  $\mathbf{M}_3$ , the predictor for the  $i^{\text{th}}$  unobserved unit is obtained by a simple back-transformation from the Box-Cox transformation:

$$\hat{y}_{i,w} = g_i(\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w) = (\hat{\lambda}_w \mathbf{x}_i' \hat{\boldsymbol{\beta}}_w + 1)^{1/\hat{\lambda}_w}, \quad (3)$$

where  $\hat{\boldsymbol{\beta}}_w$  and  $\hat{\lambda}_w$  are estimators of the model parameters. The estimation method is explained in next subsection. Equations (2) and (3) do not provide unbiased predictors for  $y_i$  under the respective underlying models. Li and Lahiri (2007) showed that if the error variance is small, that is, the model fits the data very well, the right-hand sides of equations (2) and (3) are good alternatives to the unbiased predictors. For simplicity and to reduce the computational burden, we will treat the right-hand sides of equations (2) and (3) as appropriate alternatives for the unbiased predictors. Recall that our purpose is to describe model-assisted estimators for finite population totals. The

underlying model is only used to suggest an estimator, which will be evaluated under the randomization framework. Even though the predictors of the individual  $y_i$  are biased, we can still construct design-consistent estimators for finite population totals, unlike the strictly model-based estimators proposed by Li and Lahiri (2007).

We denote GREG estimators under the three models  $\mathbf{M}_1 - \mathbf{M}_3$  by  $\hat{T}_{G-L}$ ,  $\hat{T}_{G-LOGL}$ , and  $\hat{T}_{G-BC}$ , respectively. The  $\hat{T}_{G-BC}$  estimator is different from the  $\hat{T}_{G-L}$  and  $\hat{T}_{G-LOGL}$  because the data dictates the transformation to be used.

It is possible to incorporate Box-Cox transformations on both the  $y$ -variable and  $x$ -variables. In the past, different functional forms of the Box-Cox model have been investigated. Khan and Ross (1977), Spitzer (1976), Zarembka (1968), Boylan, Cuddy and O'Muircheartaigh (1980), and others, considered a particular case of the general Box-Cox model when a common transformation parameter is assumed for the  $y$ -variable and  $x$ -variables. Gemmill (1980), Boylan, Cuddy and O'Muircheartaigh (1982), and others, applied the general Box-Cox model with different transformation parameters on  $y$ -variable and  $x$ -variables. Li and Lahiri (2007) also used the general Box-Cox model to predict the finite population total under a model-based framework. In future, this method can be extended to different functional forms of the Box-Cox transformation.

We only discuss a Box-Cox transformation of the  $y$ -variable here. This allows for a fairer comparison among the three GREG estimators.

## 2.2 Estimation of model and transformation parameters $\phi = (\beta, \lambda, \sigma^2)'$ using the pseudo-maximum likelihood (PML) method

In order to ease the estimation of  $\lambda$  using existing computational procedures, one must replace  $\mathbf{Y}^{(\lambda)}$  in the model  $\mathbf{M}_3$  by a scaled transformation  $\mathbf{Y}^{*(\lambda)}$ . For the  $i^{\text{th}}$  unit,

$$y_i^{*(\lambda)} = \begin{cases} (y_i^\lambda - 1)/\lambda \tilde{y}^{\lambda-1} & \lambda \neq 0, \\ \tilde{y} \log(y_i) & \lambda = 0, \end{cases}$$

where  $\tilde{y}$  is the geometric mean of  $y$ 's. The following calculation will be based on the new scaled model:

$$\mathbf{M}_4: \mathbf{Y}^{*(\lambda)} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}^*,$$

where  $\boldsymbol{\varepsilon}^*$  are approximately normal with mean  $\mathbf{0}$  and variance matrix  $\sigma_e^{*2}\mathbf{I}$ . Let  $\boldsymbol{\phi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma_e^{*2})'$ .

The maximum likelihood estimator (MLE) of  $\boldsymbol{\phi}^*$  maximizes the log-likelihood

$$l(\boldsymbol{\phi}^*) = \sum_i \log f(y_i; \boldsymbol{\phi}^*, \tilde{y}),$$

where

$$f(y_i; \boldsymbol{\phi}^*, \tilde{y}) = (2\pi\sigma_e^{*2})^{-1/2} \exp\left\{-(2\sigma_e^{*2})^{-1}(y_i^{*(\lambda)} - \mathbf{x}_i'\boldsymbol{\beta}^*)^2\right\} \cdot (y_i/\tilde{y})^{\lambda-1}.$$

Skinner, Holt and Smith (1989) redefines  $\boldsymbol{\phi}^*$  as the value of  $\boldsymbol{\phi}^*$  which maximizes

$$l(\boldsymbol{\phi}^*) = \sum_{i \in U} \log f(y_i, \boldsymbol{\phi}^*),$$

the sum being taken over all units in the finite population. Thus, among all possible models  $f(y_i, \boldsymbol{\phi}^*)$ , the one which "best fits" the finite population is chosen. If we choose the  $f(y_i, \boldsymbol{\phi}^*)$  family poorly, this best fit will still be poor, but our inference treats it as the target we are trying to hit with our sample data. Thus, it is important to select appropriate choices for  $f(y_i, \boldsymbol{\phi}^*)$ .

For the finite population,  $\boldsymbol{\phi}^*$  satisfies

$$\dot{l}_U(\boldsymbol{\phi}^*) = \sum_{i \in U} [\partial \log f(y_i; \boldsymbol{\phi}^*, \tilde{y}) / \partial \boldsymbol{\phi}^*] = 0,$$

where

$$\tilde{y} = \prod_{i=1}^N y_i^{1/N}.$$

For given  $\boldsymbol{\phi}^*$ , let  $\dot{l}_U(\boldsymbol{\phi}^*)$ , summation of the first derivative of the log-likelihood with respect to  $\boldsymbol{\phi}^*$ , be a finite population parameter. We take a sample, and, by approximating  $\log f(y_i; \boldsymbol{\phi}^*, \tilde{y})$  for each unit  $i$  in the sample by  $\log f(y_i; \boldsymbol{\phi}^*, \tilde{y}_w)$ , we estimate the population total,  $\dot{l}_U(\boldsymbol{\phi}^*)$ , by  $\dot{l}_s(\hat{\boldsymbol{\phi}}_{\text{PML}})$ :

$$\dot{l}_s(\hat{\boldsymbol{\phi}}_{\text{PML}}) = \sum_{i \in s} w_i [\partial \log f(y_i; \boldsymbol{\phi}^*, \tilde{y}_w) / \partial \boldsymbol{\phi}^*]_{\boldsymbol{\phi}^* = \hat{\boldsymbol{\phi}}_{\text{PML}}},$$

where

$$\tilde{y}_w = \prod_{i \in s} y_i^{w_i / \sum_{i \in s} w_i},$$

the weighted geometric mean of  $y$ 's in the sample and  $\hat{\boldsymbol{\phi}}_{\text{PML}}$  is the pseudo maximum likelihood estimator of  $\boldsymbol{\phi}^*$ , satisfying  $\dot{l}_s(\hat{\boldsymbol{\phi}}_{\text{PML}}) = 0$  (Wu and Sitter 2001). The PML estimator,  $\hat{\boldsymbol{\phi}}_{\text{PML}} = (\hat{\boldsymbol{\beta}}_w, \hat{\lambda}_w, \hat{\sigma}_{e,w}^{*2})'$ , can be obtained by a grid search method. That is, calculating and plotting the weighted log likelihood values,

$$\log L(\boldsymbol{\phi}^*) = \sum_{i \in s} w_i \log f(y_i; \boldsymbol{\phi}^*, \tilde{y}_w) \quad (4)$$

against the set of values for  $\lambda$  will locate the PML estimate,  $\hat{\lambda}_w$ , of the transformation parameter. When we evaluate the log-likelihood function at each fixed value of  $\lambda$  in the sampling context,  $\boldsymbol{\beta}^*$  and  $\sigma_e^{*2}$  are estimated by incorporating the sampling weights as:

$$\hat{\beta}_w^* = \left( \sum_{i \in S} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i \in S} w_i \mathbf{x}_i y_i^{*(\lambda)} \right),$$

$$\hat{\sigma}_{e,w}^{*2} = \sum_{i \in S} w_i (y_i^{*(\lambda)} - \mathbf{x}_i' \hat{\beta}_w^*)^2 / \sum_{i \in S} w_i.$$

Since model  $\mathbf{M}_3$  is of the interest, converting  $\hat{\Phi}_{\text{PML}}^*$  that maximizes (4) back to  $\hat{\Phi}_{\text{PML}}$  in the model  $\mathbf{M}_3$  is necessary and  $\hat{\beta}_w = \hat{\gamma}_w^{-1} \hat{\beta}_w^*$ ,  $\hat{\sigma}_{e,w}^2 = \hat{\gamma}_w^{2(\hat{\lambda}_w-1)} \hat{\sigma}_{e,w}^{*2}$ .

Since we plan to use a design-consistent estimator for the population total, it is reasonable to question why we need the sampling weights in estimating the model parameters. Sverchkov and Pfeffermann (2004) argue that incorporating weights can produce better estimates of model parameters when the model is correct in the population but wrong in the sample. That can happen when the probabilities or selections are correlated within the  $\varepsilon_i$ .

Let  $\mathbf{B}^*$  and  $\Lambda$  be the first two components of the full-sample solution to the maximization of  $\sum_U \log f(y_i, \hat{\Phi}^*)$ , and  $\mathbf{B} = \hat{\gamma}^{\Lambda-1} \mathbf{B}^*$ . This allows us to define  $\theta_N = (\mathbf{B}, \Lambda)$  as finite population parameters, irrespective of the validity of the model.

It can be shown that  $\hat{\theta}_w = (\hat{\beta}_w, \hat{\lambda}_w)$  is a design-consistent estimator of  $\theta_N$ , that is,

$$\hat{\beta}_w \rightarrow \mathbf{B} \text{ in probability and } \hat{\lambda}_w \rightarrow \Lambda \text{ in probability}$$

under certain regularity conditions using arguments similar to Binder (1983), Wu (1999) and Wu and Sitter (2001). Here the probabilistic convergence is with respect to the sampling design.

### 2.3 Consistency property of $\hat{T}_{G\_BC}$ estimator

It is well-known that the  $\hat{T}_{G\_L}$  estimator has the desirable property of design-consistency under mild conditions (Särndal *et al.* 1992). This means that the relative difference between the estimator and what it estimates will converge in probability to 0 as the sample grows arbitrarily large whether or not the working model on which it is based holds. That property can be maintained for the  $\hat{T}_{G\_BC}$  estimator.

Define

$$\hat{T}_{D\_G\_BC} = \sum_{i \in U} g_i(\theta_N) + \sum_{i \in S} (y_i - g_i(\theta_N)) / \pi_i,$$

and

$$\hat{T}_{G\_BC} = \sum_{i \in U} g_i(\hat{\theta}_w) + \sum_{i \in S} (y_i - g_i(\hat{\theta}_w)) / \pi_i.$$

**Theorem:** Under the following assumptions, the Box-Cox-based GREG estimator  $\hat{T}_{G\_BC}$  is design consistent for  $T$ , in the sense that  $N^{-1}(\hat{T}_{G\_BC} - T) = O_p(1/\sqrt{n})$ . Furthermore, the asymptotic variance of  $\hat{T}_{G\_BC}$  is given by

$$AV_d(\hat{T}_{G\_BC}) =$$

$$\sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) (y_i - g_i(\theta_N)) (y_j - g_j(\theta_N)) / (\pi_i \pi_j),$$

which can be estimated by

$$\hat{V}_d(\hat{T}_{G\_BC}) =$$

$$\sum_{i \in S} \sum_{j \in S} ((\pi_{ij} - \pi_i \pi_j) / \pi_{ij}) (y_i - g_i(\hat{\theta}_w)) (y_j - g_j(\hat{\theta}_w)) / (\pi_i \pi_j). \quad (5)$$

Assumption 1:  $\hat{\theta}_w = \theta_N + O_p(1/\sqrt{n})$ ;

Assumption 2: For each  $\mathbf{x}_i$ ,  $\partial g_i(\mathbf{t}) / \partial \mathbf{t}$  is continuous in  $\mathbf{t}$  and  $|\partial g_i(\mathbf{t}) / \partial \mathbf{t}| \leq h(\theta)$  for  $\mathbf{t}$  in a neighbourhood of  $\theta$ ;

Assumption 3: The Horvitz-Thompson estimators with the basic design weights for certain population totals are asymptotically normally distributed.

Assumption 4: For each  $\mathbf{x}_i$ , the second derivative of  $g_i(\mathbf{t})$  with respect to  $\mathbf{t}$  is continuous and bounded in the neighborhood of  $\theta$ .

Proof: (see Appendix).

The proposed variance estimator in equation (5) is based on large sample approximations. For a given nominal level  $1 - \alpha$ , the usual confidence interval based on the normal approximation for the variance estimator gives approximately  $100(1 - \alpha)\%$  coverage rate in repeated large samples. Unfortunately, in some cases, it has been observed that the coverage properties of this type of variance estimator can be poor for some choices of the assisted model for the  $\hat{T}_{G\_L}$  estimators (Särndal 1982; Särndal, Swensson and Wretman 1989; *etc.*). Theoretical and empirical studies on the coverage property of the proposed variance estimator need further investigation.

The  $\hat{T}_{G\_BC}$  estimator is design-consistent for the finite population total  $T$  under the randomization approach, and the Box-Cox technique allows a reasonable transformation on the dependent variable to be automatically determined by the data from a large family of functions, and hence increased efficiency can be achieved.

### 3. A simulation study

The purpose of this simulation study is to evaluate the performance of different GREG estimators for a finite population total. In this simulation exercise, a finite population from the Australian Agricultural and Grazing Industries Survey (AAGIS) is generated. This survey data contains information on the number of cattle ( $y$ ) and farm area ( $x$ ) for each of the 431 farms.



We consider a finite population of size  $N=4,000$ , generated from the following model:

$$\mathbf{M}_5: y_i^{(\lambda)} = (y_i^\lambda - 1)/\lambda = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

where  $\varepsilon_i$ 's are independent with approximate  $N(0, \sigma^2)$ , and  $x_i$  is the logarithm of a value generated from an exponential distribution with mean  $\mu_x$  and standard error  $\sigma_x$ . In order to mimic a true situation, we choose  $\lambda = 0.1$ ,  $\beta_0 = 4.20$ , and  $\beta_1 = 2.66$  which are the estimates obtained by fitting the real survey data to the model  $\mathbf{M}_3$ . We set  $\mu_x = 1,040$ ,  $\sigma_x = 1,000$  to ensure  $y_i > 0$  for almost all unit  $i$ . Strictly speaking, we have a truncated normal distribution of  $y$ , since all negative values of  $y$  generated are discarded. The effect of this is negligible since less than 0.1% of the generated  $y$  values need to be discarded. The same phenomenon was found by Taylor (1986).

Simulation is based on repeated sampling from the generated finite population. Two sampling designs are investigated: simple random sampling (SRS) and stratified SRS (STSRs). When a sample is selected by STSRs, unequal selection probabilities among different strata are applied. We define two strata using the boundary value: median of  $y$  values in the finite population. For stratum  $h$  of size  $N_h$ , a simple random sample of size  $n_h$  is selected. Define  $p_1$  and  $p_2$  selection probabilities for stratum 1 and stratum 2, respectively. We specify  $p_1 = 2 \times p_2$ . For fixed sample size  $n$ ,  $n_1 = N_1 \times p_1$ , and  $n_2 = N_2 \times p_2$ .

We are interested in estimating the finite population total

$$T = \sum_{i \in U} y_i.$$

In this simulation study, we study the performances of  $\hat{T}_{G\_BC}$  estimator,  $\hat{T}_{G\_L}$ ,  $\hat{T}_{G\_LOGL}$ , along with the design-based Horvitz-Thompson estimator ( $\hat{T}_D$ ), where the subscripts “-L”, “-LOGL”, and “-BC” denote the underlying linear model, log-linear model, and Box-Cox model, respectively.

One thousand samples are selected from the simulated finite population for each of the sample size  $n \in (30, 80, 150)$ . Four estimators are produced for each selected sample. Estimator of the finite population transformation parameter  $\Lambda$  is also produced for each sample. For the purpose of comparison, two methods are used to estimate  $\Lambda$ . Let  $\hat{\lambda}(\hat{\lambda}_u)$  be the OLS/ML (PML) estimators of  $\Lambda$ . Over all the 1,000 samples, we compute the empirical percentage relative biases (RelBias) and root mean square errors (rmse) to evaluate these estimators using the following formulae:

$$\text{RelBias} = B^{-1} \sum_{b=1}^B (\hat{\omega}_b - \omega) / \omega,$$

and

$$\text{rmse} = \sqrt{B^{-1} \sum_{b=1}^B (\hat{\omega}_b - \omega)^2},$$

where  $B$  is the number of the replications in the Monte Carlo simulation and  $\hat{\omega}$  represents an arbitrary estimates of the finite population parameter  $\omega$ .

#### 4. Results

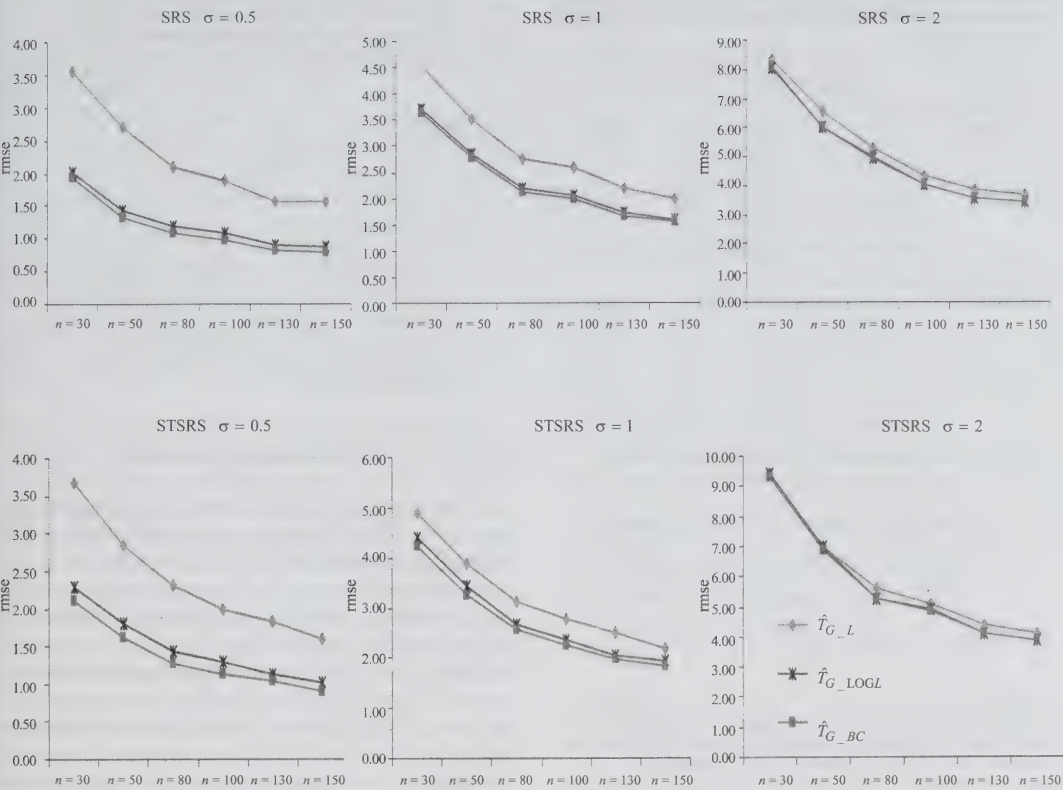
In Table 1 we present the RelBias and rmse of four estimators using different sampling designs with varying sample sizes when  $\sigma = 0.5$ . All four estimators give RelBias close to zero [maximum of the absolute values of RelBias (|RelBias|) in Table 1 is less than 0.01]. Among them,  $\hat{T}_{G\_BC}$  has the smallest |RelBias| and rmse over different sampling sizes and sampling designs. Therefore, the Box-Cox technique protects  $\hat{T}_{G\_BC}$ , which achieves improvement in efficiency compared to other GREG estimators.

The RelBias and rmse under the same conditions when  $\sigma = 1$  and  $\sigma = 2$  are also investigated. Figure 1 presents the rmse for the three GREG estimators ( $\hat{T}_{G\_L}$ ,  $\hat{T}_{G\_LOGL}$ , and  $\hat{T}_{G\_BC}$ ) using different sampling designs. We note that  $\hat{T}_{G\_BC}$  consistently has the smallest rmse when  $\sigma = 0.5$  and 1. Thus, a robust model chosen by the Box-Cox method reduces the rmse, especially when the model is appropriate with small  $\sigma$ .

**Table 1**  
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes ( $\lambda = 0.1$ )

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_LOGL}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	4.37	-9.82	5.46	3.62
	$n = 80$	1.43	-3.67	1.24	0.65
	$n = 150$	-2.60	-1.24	1.22	0.53
rmse( $\times 10^7$ )	$n = 30$	7.17	3.54	2.02	1.94
	$n = 80$	4.26	2.09	1.18	1.09
	$n = 150$	3.20	1.58	0.88	0.79
Stratified simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	6.01	-5.82	3.84	1.92
	$n = 80$	9.93	-1.01	3.04	0.98
	$n = 150$	1.75	-1.85	1.06	0.42
rmse( $\times 10^7$ )	$n = 30$	5.63	3.67	2.29	2.11
	$n = 80$	3.51	2.31	1.43	1.28
	$n = 150$	2.49	1.59	1.01	0.90

- <sup>1</sup>  $\hat{T}_D$ : the design-based Horvitz-Thompson estimator;  
<sup>2</sup>  $\hat{T}_{G\_L}$ : the GREG estimator with the underlying linear model;  
<sup>3</sup>  $\hat{T}_{G\_LOGL}$ : the GREG estimator with the underlying log-linear model;  
<sup>4</sup>  $\hat{T}_{G\_BC}$ : the GREG estimator with the underlying Box-Cox model.



**Figure 1** Comparison of root mean square error of  $\hat{T}_{G,L}$ ,  $\hat{T}_{G,LOGL}$ , and  $\hat{T}_{G,BC}$  with varying sampling designs, sample sizes, and standard deviations

We may also be interested in how  $\hat{T}_{G,L}$  performs compared to  $\hat{T}_{G,BC}$  when  $\lambda = 1$ , the situation favoring  $\hat{T}_{G,L}$ . From Table 2, we can see that  $\hat{T}_{G,BC}$  is quite comparable to  $\hat{T}_{G,L}$ , especially when sample size is large. This result implies we don't have much loss in RelBias and rmse by using  $\hat{T}_{G,BC}$  when  $\hat{T}_{G,L}$  should be used.

In order to better assess the robustness and the improvement in efficiency of the  $\hat{T}_{G,BC}$  versus  $\hat{T}_{G,L}$  and  $\hat{T}_{G,LOGL}$ , a finite population of size  $N = 4,000$  is generated from the model not in agreement with the Box-Cox model  $\mathbf{M}_5$ , but

$$\mathbf{M}_6: y_i^{(\lambda)} = \beta_0 + \beta_1 x_i + z_i + \varepsilon_i,$$

where  $z_i = x_i^2$ . The same  $x_i$ 's and parameter values as specified in Section 3 are used to generate  $y_i$ 's. The same four estimators are studied based on the new finite population  $\{(x_i, y_i)\}$ 's for  $i = 1, \dots, 4,000$ . This is the situation not ideal for any of the GREG estimators. The

results are shown in Table 3. The advantage of using  $\hat{T}_{G,BC}$  is obvious in terms of the RelBias and rmse.

Table 4 presents the RelBias and rmse of  $\hat{\lambda}$  and  $\hat{\lambda}_w$  for STSRs sampling with varying sample sizes and  $\sigma$ 's based on population values generated from  $\mathbf{M}_5$ . Since we stratified on the  $y$ -variable, using the weights should have an impact, at least on the bias of the parameter estimate. When  $\sigma$  is small, that is, when the simulated data are well fitted to the assumed model,  $\hat{\lambda}_w$  gives RelBias closer to zero, but  $\hat{\lambda}_w$  and  $\hat{\lambda}$  perform equally well in terms of the rmse. When  $\sigma$  is large, however,  $\hat{\lambda}_w$  consistently gives smaller absolute values of RelBias and rmse, as compared to  $\hat{\lambda}$ , although the rmse's remain close. Indeed, when it come to estimating  $T$ , neither approach has the advantage in terms of empirical bias or root mean squared error no matter the sample size or the setting for  $\sigma$  (not shown). This may be because the estimator for  $T$  is model-biased no matter how well  $\lambda$  is estimated.

**Table 2**  
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes ( $\lambda = 1$ )

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_LOGL}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	0.31	0.16	0.63	0.24
	$n = 80$	-0.37	-0.09	0.03	-0.09
	$n = 150$	-0.25	0.03	0.10	0.02
rmse( $\times 10^7$ )	$n = 30$	20.91	3.51	3.98	3.63
	$n = 80$	12.22	2.12	2.36	2.13
	$n = 150$	8.98	1.57	1.75	1.57
Stratified simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	0.51	0.27	0.56	0.22
	$n = 80$	3.97	-0.11	-0.11	-0.13
	$n = 150$	-0.23	0.04	0.06	0.04
rmse( $\times 10^7$ )	$n = 30$	12.54	3.79	4.20	3.91
	$n = 80$	8.39	2.27	2.61	2.29
	$n = 150$	5.48	1.67	1.90	1.67

<sup>1</sup>  $\hat{T}_D$ : the design-based Horvitz-Thompson estimator;  
<sup>2</sup>  $\hat{T}_{G\_L}$ : the GREG estimator with the underlying linear model;  
<sup>3</sup>  $\hat{T}_{G\_LOGL}$ : the GREG estimator with the underlying log-linear model;  
<sup>4</sup>  $\hat{T}_{G\_BC}$ : the GREG estimator with the underlying Box-Cox model.

**Table 3**  
Relative biases and root mean square errors of the four estimators using different sampling designs with varying sample sizes ( $y$  values generated from a model  $M_6$ )

		$\hat{T}_D^1$	$\hat{T}_{G\_L}^2$	$\hat{T}_{G\_LOGL}^3$	$\hat{T}_{G\_BC}^4$
Simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	-16.89	-54.56	28.81	-2.17
	$n = 80$	-5.65	-23.15	11.76	-1.43
	$n = 150$	-13.78	-13.78	10.73	-0.76
rmse( $\times 10^{11}$ )	$n = 30$	30.08	24.68	12.11	2.87
	$n = 80$	17.95	13.85	7.25	1.78
	$n = 150$	13.60	10.29	5.50	1.33
Stratified simple random sampling					
RelBias( $\times 10^{-3}$ )	$n = 30$	1.11	-36.13	31.95	-7.87
	$n = 80$	5.59	-18.13	13.11	-3.26
	$n = 150$	-2.79	-7.30	7.33	-1.43
rmse( $\times 10^{11}$ )	$n = 30$	34.10	27.15	14.19	4.37
	$n = 80$	19.61	15.93	8.49	2.61
	$n = 150$	14.60	12.19	6.62	2.02

<sup>1</sup>  $\hat{T}_D$ : the design-based Horvitz-Thompson estimator;  
<sup>2</sup>  $\hat{T}_{G\_L}$ : the GREG estimator with the underlying linear model;  
<sup>3</sup>  $\hat{T}_{G\_LOGL}$ : the GREG estimator with the underlying log-linear model;  
<sup>4</sup>  $\hat{T}_{G\_BC}$ : the GREG estimator with the underlying Box-Cox model.

**Table 4**  
Relative biases and root mean square error of  $\hat{\lambda}^1$  and  $\hat{\lambda}_w^2$  for STSRS sampling with varying sample sizes and standard deviations.

	$\sigma = 2$		$\sigma = 1$		$\sigma = 0.5$	
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}$	$\hat{\lambda}_w$
Relative biases						
$n = 30$	-0.58	0.13	-0.28	0.10	-0.16	-0.01
	-0.42	0.14	-0.19	0.11	-0.13	-0.02
	-0.39	0.10	-0.16	0.09	-0.10	-0.01
Root mean square error						
$n = 30$	0.14	0.12	0.11	0.11	0.07	0.07
	0.08	0.07	0.06	0.06	0.04	0.04
	0.06	0.05	0.05	0.04	0.03	0.03

<sup>1</sup> Estimator  $\hat{\lambda}$  is obtained using ordinary least square method/maximum likelihood method;  
<sup>2</sup> Estimator  $\hat{\lambda}_w$  is obtained by pseudo-maximum likelihood method.

5. Concluding remarks

In this article, we have proposed a generalized regression estimator of a finite population total based on the Box-Cox transformation technique under a general unequal probability sampling design. The proposed estimator, being design-consistent, maintains the robustness property of GREG even if the underlying model fails. In many situations, some version of the model in  $M_3$  will at least provide a useful approximation of dependent-variable behavior. The Box-Cox technique allows a reasonable-fitting transformation on the dependent variable to be automatically determined by the data. The robustness and efficiency of the proposed estimator were evaluated analytically and via Monte Carlo simulations.

When comparing a GREG based on an underlying linear model ( $\hat{T}_{G\_L}$ ) to one based on a Box-Cox model ( $\hat{T}_{G\_BC}$ ), we should remember that  $\hat{T}_{G\_L}$  doesn't require complete auxiliary information. Moreover, it can produce a single set of weights usable for all variables of interest, unlike the  $\hat{T}_{G\_BC}$ . To achieve higher efficiency, however, both estimators usually require different weights for different variables of interest, because each study variable is best fit by its own working model. The  $\hat{T}_{G\_BC}$  can provide even more efficiency than the  $\hat{T}_{G\_L}$  but at the cost of requiring complete information about the  $x$ -variables. Such information, although rarely available in North American household surveys, is often available in business surveys.

Surveys are rarely conducted to measure a single variable of interest. The question is how to estimate the finite totals for mutually exclusive and exhaustive subpopulations such that those estimates will add up to the estimate of the finite total for the entire population. We need to take special care of this problem since Box-Cox estimators are not linear in  $y$ . Such estimates can be obtained using a standard



benchmarking tool outlined in Li and Lahiri (2007). Another approach is to use model calibration (Wu and Sitter 2001). Treating the predictions  $\hat{y}_{i,w}$  from equation (3) as the auxiliary variable into a design-consistent linear regression estimator, such as

$$\hat{T}_{G\_BC} = \sum_{i \in S} y_i / \pi_i + \left( \sum_{i \in U} \mathbf{z}'_i - \sum_{i \in S} \mathbf{z}'_i / \pi_i \right) \mathbf{d},$$

where

$$\mathbf{d} = \left( \sum_{i \in S} w_i \mathbf{z}_i \mathbf{z}'_i \right)^{-1} \left( \sum_{i \in S} w_i \mathbf{z}_i y_i \right), \text{ and } \mathbf{z}_i = (1 \ \hat{y}_{i,w})',$$

will produce a set of calibration weights

$$(w_{i,c} = [1 + (\sum_U \mathbf{z}'_i - \sum_S \mathbf{z}'_i / \pi_i) (\sum_S w_j \mathbf{z}_j \mathbf{z}'_j)^{-1} \mathbf{z}_i] w_i)$$

that can be used generally. Moreover, we can, in principle, incorporate more than one set of predictors as auxiliary variables, either for different variables of interest or the same variable of interest broken into several subpopulations. This is a fruitful area for future research.

Several other extensions of our current method merit further exploration. We did not consider here the possibility of unit model errors having a complex correlation structure. Although a design-consistent estimator obtains using our methods when such a structure exists but is ignored, the efficiency of the estimator likely suffers. It will be interesting to investigate whether allowing certain correlation structures in the data can make the estimation procedure more efficient.

In this article, we only transformed  $y$ -variable by the Box-Cox technique. In future, this method can be extended to different functional forms of the Box-Cox transformation.

The variance estimator of the  $\hat{T}_{G\_BC}$  estimator proposed in this article is based on large sample approximation. Some studies showed poor performance for this type of variance estimator for some choices of the assisted models for the  $\hat{T}_{G\_L}$ . Theoretical and empirical studies on the coverage property of the proposed variance estimator need further investigation.

## Acknowledgements

I wish to thank the Editor and three referees for a number of constructive suggestions which led to substantial improvement of the original manuscript. I thank Dr. Phil Kott for his generous line editing of the text and Dr. Alan Dorfman for supplying the survey data used in the paper. The author is grateful to her Ph.D. advisor Professor Partha Lahiri, supervisory committee members Professors Katherine Abraham, Wolfgang Jank, Fritz Scheuren,

Paul Smith and Mr. Paul D. Williams for their generous guidance throughout her dissertation research.

## Appendix

Write

$$\hat{T}_{G\_BC} = \sum_{i \in S} y_i / \pi_i + \left( \sum_{i \in U} g_i(\hat{\theta}_w) - \sum_{i \in S} g_i(\hat{\theta}_w) / \pi_i \right),$$

by Taylor Series expansion, with assumption (2) we have

$$g_i(\hat{\theta}_w) = g_i(\theta_N) + (\partial g_i(\mathbf{t}) / \partial \mathbf{t})|_{\mathbf{t}=\theta^*} (\hat{\theta}_w - \theta_N),$$

where  $\theta^* \in (\theta_N, \hat{\theta}_w)$  or  $(\hat{\theta}_w, \theta_N)$ , and  $(\partial g_i(\mathbf{t}) / \partial \mathbf{t})$  is a row vector.

By assumptions (1) and (2),

$$N^{-1} \sum_{i \in U} g_i(\hat{\theta}_w) = N^{-1} \sum_{i \in U} g_i(\theta_N) + O_p(1/\sqrt{n})$$

$$N^{-1} \sum_{i \in S} \pi_i^{-1} g_i(\hat{\theta}_w) = N^{-1} \sum_{i \in S} \pi_i^{-1} g_i(\theta_N) + O_p(1/\sqrt{n}).$$

Also note that by assumption (3),

$$N^{-1} \sum_{i \in U} g_i(\theta_N) = N^{-1} \sum_{i \in S} \pi_i^{-1} g_i(\theta_N) + O_p(1/\sqrt{n}).$$

Therefore,

$$N^{-1} \left( \sum_{i \in U} g_i(\hat{\theta}_w) - \sum_{i \in S} \pi_i^{-1} g_i(\hat{\theta}_w) \right) = O_p(1/\sqrt{n}).$$

Also, by assumption (3),

$$N^{-1} \left( \sum_{i \in U} y_i - \sum_{i \in S} \pi_i^{-1} y_i \right) = O_p(1/\sqrt{n}).$$

Therefore,  $N^{-1}(\hat{T}_{G\_BC} - T) = O_p(1/\sqrt{n})$ , i.e.,  $\hat{T}_{G\_BC}$  converges in probability to  $T$  with the order of  $O_p(1/\sqrt{n})$ .

In addition, with assumption (4), a second-order Taylor series approximation to  $g_i(\hat{\theta}_w)$  can be expanded as:

$$g_i(\hat{\theta}_w) = g_i(\theta_N) + (\partial g_i(\mathbf{t}) / \partial \mathbf{t})'|_{\mathbf{t}=\theta_N} (\hat{\theta}_w - \theta_N) + (\hat{\theta}_w - \theta_N)' (\partial^2 g_i(\mathbf{t}) / (\partial \mathbf{t} \partial \mathbf{t}')|_{\mathbf{t}=\theta^*} (\hat{\theta}_w - \theta_N),$$

where  $\theta^* \in (\theta_N, \hat{\theta}_w)$  or  $(\hat{\theta}_w, \theta_N)$ . It follows from assumptions (1) and (4) that

$$N^{-1} \sum_{i \in U} g_i(\hat{\theta}_w) = N^{-1} \sum_{i \in U} g_i(\theta_N) + N^{-1} \sum_{i \in U} (\partial g_i(\mathbf{t}) / \partial \mathbf{t})'|_{\mathbf{t}=\theta_N} (\hat{\theta}_w - \theta_N) + O_p(n^{-1}),$$

$$N^{-1} \sum_{i \in S} \pi_i^{-1} g_i(\hat{\theta}_w) = N^{-1} \sum_{i \in S} \pi_i^{-1} g_i(\theta_N) \\ + N^{-1} \sum_{i \in S} \pi_i^{-1} (\partial g_i(\mathbf{t}) / \partial \mathbf{t}')|_{\mathbf{t}=\theta_N} (\hat{\theta}_w - \theta_N) + O_p(n^{-1}).$$

By assumptions (1) and (3),  $\hat{\theta}_w = \theta_N + O_p(1/\sqrt{n})$  and

$$N^{-1} \sum_{i \in S} \pi_i^{-1} \partial g_i(\mathbf{t}) / \partial \mathbf{t}'|_{\mathbf{t}=\theta_N} = N^{-1} \sum_{i \in U} \partial g_i(\mathbf{t}) / \partial \mathbf{t}'|_{\mathbf{t}=\theta_N} \\ + O_p(1/\sqrt{n}).$$

Hence,

$$N^{-1} \left( \sum_{i \in U} g_i(\hat{\theta}_w) - \sum_{i \in S} \pi_i^{-1} g_i(\hat{\theta}_w) \right) = \\ N^{-1} \left( \sum_{i \in U} g_i(\theta_N) - \sum_{i \in S} \pi_i^{-1} g_i(\theta_N) \right) + O_p(n^{-1}).$$

Therefore,

$$\hat{T}_{G\_BC} = \sum_{i \in S} y_i / \pi_i \\ + \left( \sum_{i \in U} g_i(\theta_N) - \sum_{i \in S} \pi_i^{-1} g_i(\theta_N) \right) + O_p(N/n).$$

The asymptotic design-variance of  $\hat{T}_{AG}$  is:

$$AV_d(\hat{T}_{G\_BC}) \\ \approx \sum_{i \in U} \sum_{j \in U} (\pi_{ij} - \pi_i \pi_j) (y_i - g_i(\theta_N)) (y_j - g_j(\theta_N)) / (\pi_i \pi_j),$$

which can be estimated by

$$\hat{V}_d(\hat{T}_{G\_BC}) \\ \approx \sum_{i \in S} \sum_{j \in S} (\pi_{ij} - \pi_i \pi_j) / \pi_{ij} (y_i - g_i(\hat{\theta}_w)) (y_j - g_j(\hat{\theta}_w)) / (\pi_i \pi_j).$$

## References

- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *International Statistical Review*, 51, 279-92.
- Bickel, P.J., and Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296-311.
- Box, G.E., and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Boylan, T.A., Cuddy, M.P. and O'Muircheartaigh, I.G. (1980). The functional form of the aggregate import demand equation: A comparison of three european economies. *Journal of international economics*, 10, 561-566.
- Boylan, T.A., Cuddy, M.P. and O'Muircheartaigh, I.G. (1982). Import demand equations: An application of a generalized Box-Cox methodology. *International Statistical Review*, 50, 103-112.
- Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.
- Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- Carroll, R.J., and Ruppert, D. (1988). *Transformations and weighting in Regression*. London: Chapman and Hall.
- Cassel, C.M., Särndal, C.-E. and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- Chambers, R.L., and Dorfman, A.H. (2003). Transformed variables in survey sampling. Technical Report.
- Chen, G., and Chen, J. (1996). A transformation method for finite population sampling calibrated with empirical likelihood. *Survey Methodology*, 22, 139-146.
- Davison, C.W., Arnade, C.A. and Hallahan, C.B. (1989). Box-Cox estimation of U.S. soyabean exports. *Journal of Agricultural Economics Research*, 41, 8-15.
- Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Esteveao, V., Hidirolou, M. and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-25.
- Fuller, W.A., Loughin, M. and Baker, H. (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey. *Survey Methodology*, 20, 75-85.
- Gemmil, G. (1980). Using the Box-Cox Form for Estimating Demand: A Comment. *The review of economics and statistics*, 62, 147-148.
- Gurka, M. (2004). The Box-Cox transformation in the general linear mixed model for longitudinal data. Ph.D. Dissertation.
- Gurka, M. (2006). Expanding the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society, Series A*, 169, 273-288.
- Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association*, 47, 663-685.
- Jayasuriya, B., and Valliant, R. (1996). An application of restricted regression estimation to post-stratification in a household survey. *Survey Methodology*, 22, 127-137.
- Jiang, J., and Lahiri, P. (2006). Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.

- John, N.R., and Draper, J.A. (1980). An alternative family of transformations. *Applied Statistics*, 29, 190–197.
- Karlberg, F. (2000). Survey estimation for highly skewed population in the presence of zeroes. *Journal of Official Statistics*, 16, 229–241.
- Khan, M.S., and Ross, K.Z. (1977). The functional form of aggregate import demand. *Journal of International Economics*, 7, 149–160.
- Korn, E., and Graubard, B. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24, 193–201.
- Li, Y., and Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664–673.
- Miner, A.G. (1982). The contribution of weather and technology to U.S. soybean yield. *Unpublished Dissertation*. University of Minnesota.
- Montanari, G.E., and Ranalli, M.G. (2003). Nonparametric model calibration estimation in survey sampling. Manuscript.
- Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.
- Newman, P. (1977). Malaria and Mortality. *Journal of the American Statistical Association*, 72, 257–263.
- Royall, R.M., and Cumberland, W.G. (1981). The finite population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924–930.
- Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169–178.
- Särndal, C.-E. (1980). On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639–650.
- Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Plan Inference*, 7, 155–170.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527–537.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.
- Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *Journal of the Royal Statistical Society, Series B*, 33, 307–311.
- Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys*. Chichester: Wiley.
- Spitzer, J.J. (1976). The demand for money, the liquidity trap and functional forms. *The International Economics Review*, 17, 220–227.
- Sverchkov, M., and Pfeiffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79–92.
- Taylor, M.J.G. (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114–118.
- Tukey, J.W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics*, 28, 601–632.
- Wu, C. (1999). The effective use of complete auxiliary information from survey data. Ph.D. dissertation, Simon Fraser University, Canada.
- Wu, C., and Sitter, R.R. (2001). A model-calibration to using complete auxiliary information from survey data, *Journal of the American Statistical Association*, 96, 185–193.
- Zarembka, P. (1968). Functional form in the demand for money. *Journal of the American Statistical Association*, 63, 502–511.
- Zheng, H., and Little, R.J.A. (2004). Penalized spline nonparametric mixed models for inference about a finite population mean from two-stage samples. *Survey Methodology*, 30, 209–218.





# The BACON-EEM algorithm for multivariate outlier detection in incomplete survey data

Cédric Béguin and Beat Hulliger<sup>1</sup>

## Abstract

With complete multivariate data the BACON algorithm (Billor, Hadi and Vellemann 2000) yields a robust estimate of the covariance matrix. The corresponding Mahalanobis distance may be used for multivariate outlier detection. When items are missing the EM algorithm is a convenient way to estimate the covariance matrix at each iteration step of the BACON algorithm. In finite population sampling the EM algorithm must be enhanced to estimate the covariance matrix of the population rather than of the sample. A version of the EM algorithm for survey data following a multivariate normal model, the EEM algorithm (Estimated Expectation Maximization), is proposed. The combination of the two algorithms, the BACON-EEM algorithm, is applied to two datasets and compared with alternative methods.

**Key Words:** Forward search method; Outlier detection; Multivariate data; Missing value; Sampling; Robustness; EM-algorithm.

## 1. Introduction

The problem underlying the methods presented in this article is a sample survey on quantitative data like sales of different products where missing values and outliers occur. Often in the editing phase of the survey outliers are detected by inspection of individual questionnaires or by univariate outlier detection methods. However, there are few systematic methods which allow multivariate outlier detection in incomplete survey data.

Outlier detection is an important aspect of statistical data editing. Undetected outliers may have a large and undesirable impact on survey results. Most existing outlier-detection methods are designed for complete univariate or bivariate data. However, real outliers in survey data are often multivariate in nature. The problem of outliers becomes much more difficult in three or more dimensions than in one dimension or two. While an outlier can only be very small or very large in one dimension (at least for unimodal distributions) in higher dimensions the issue of the “direction” of the outlier becomes more and more important. Outliers may be quite close to the bulk of the data or to a model if the distance is measured in a Euclidean metric because this metric only checks the axis directions. However, if a metric appropriate to the correlation structure of the bulk of the data is used the outlier may be far away. Thus in higher dimensions the form of the point cloud of the bulk of the data must be well reflected in the metric used to detect outliers.

Outlier detection needs a model for the bulk of the data to be able to distinguish observations which are not fitted well by the model. Thus outlier detection is inherently tied to models and their robust estimation. In a sampling context

the model should be appropriate for the bulk of the population and not only for the bulk of the sample. Therefore, the sample design should be taken into account when detecting outliers in sample survey data. The discussion on the role of sampling weights is taken up again in Sections 1 and 5.

Survey data often contains missing values. Outlier detection with missing items must estimate the model for the bulk of the data taking into account the missingness. This estimation under missing values will be based on the relationship among the observed and missing variables. The relationship must be modeled robustly to protect it from outliers. If an observation would be classified as an outlier based on complete information but the values causing the outlyingness are missing then the outlier will not show up compared with a robust model. Therefore it will be difficult to detect an outlier which is outlying only in its missing values. This is analogue to the conception of missingness at random (MAR) (Little and Rubin 1987): We need information in the observed values to infer that an observation is an outlier. We may call this situation “outlying at random”. We can formalize it by stating that the outlier mechanism does not depend on unobserved data, which includes unobserved true values of the outlier in case the outlier is an error. However, for outlier detection this condition is too strict because we may be able to detect outliers in observed values even if the mechanism depends on unobserved values. This is possible because the model must hold for the bulk of the data only and not for the outliers. If the observed values of the outlier deviate enough from the model the outlier will be detected. However, when it comes to imputation of true values for nominated outliers we are in the same situation as for missing values. If the

1. Cédric Béguin, University of Neuchâtel, 2010 Neuchâtel, Switzerland; Beat Hulliger, University of Applied Sciences Northwestern Switzerland, 4600 Olten, Switzerland. E-mail: beat.hulliger@fhnw.ch.

outlier mechanism, conditionally on the observed values, still depends on the true unobserved values of the outlier we cannot estimate a model for the unobserved values. In this article we use imputation only as an ad hoc device for better outlier detection. Nevertheless we assume that, conditionally on the observed data, the non-response mechanism and the outlier-mechanism are independent and that both mechanisms do not depend on unobserved data.

In a workpackage of the EUREDIT project on "The development and evaluation of new methods for editing and imputation" (EUREDIT 2003) the authors developed outlier detection methods which cope with this difficult set-up: multivariate incomplete sample survey data. Two of these methods, Transformed Rank Correlations and the Epidemic Algorithm are presented in (Béguin and Hulliger 2004). The third method, BACON-EEM, is presented here.

In this article we concentrate on outlier detection. The scenario we have in mind is that once an outlier is detected, either it may be checked and treated manually or it may be treated by imputation. Robust estimation would replace both detection and imputation but is less adapted to the practice of official statistics. We do not distinguish between representative and non-representative outliers (Chambers 1986) since both types of outliers have to be detected, though they may have to be treated differently.

For complete data the existing multivariate methods can be classified into two major families. Many methods suppose that the data follow some elliptical distribution and try to estimate robustly the center and the covariance matrix. Then they use a corresponding Mahalanobis distance to detect outliers. The second class of methods does not rely on a distributional assumption but uses some measure of data-depth (see Liu, Parelius and Singh 1999, for a review) to be used as an outlyingness measure. The second family is at first sight more appealing, but, unfortunately, it often fails to yield methods computationally feasible with large datasets.

Many robust estimators of the covariance matrix have been reported in the literature. M-estimators (Huber 1981; Maronna 1976) have the advantage of being relatively simple to compute with a straightforward iteration from a good starting point (Rocke and Woodruff 1993). But their breakdown point - *i.e.*, the smallest fraction of the data whose arbitrary modification can carry an estimator beyond all bounds - is at most  $1/(p+1)$  where  $p$  is the dimension of the data (Donoho 1982; Maronna 1976; Stahel 1981). This handicap is important when dealing with data from official statistics, which is often high dimensional. Many other affine equivariant robust estimators, *i.e.*, estimators which transform coherently when the data is transformed linearly, were studied by (Donoho 1982) but all have breakdown points of at most  $1/(p+1)$ . Other approaches ended up with affine equivariant high breakdown point estimators, *e.g.* the

Stahel-Donoho (SD) estimator (Stahel 1981; Donoho 1982) or the Minimum Covariance Determinant (MCD) estimators (Rousseeuw 1985; Rousseeuw and Leroy 1987), but had the disadvantage of being computationally expensive. An approach of Gnanadesikan and Kettenring, using a componentwise construction of the covariance matrix, sacrificed affine equivariance but gained simplicity and speed. This approach has been re-actualized in (Maronna and Zamar 2002) and in one of the methods presented in (Béguin and Hulliger 2004), called Transformed Rank Correlations (TRC). TRC calculates an initial matrix of bivariate Spearman Rank correlations. To ensure positive definiteness of the covariance matrix the data is transformed into the space of eigenvectors of the initial matrix. The coordinatewise medians and median absolute deviations in this new space are then backtransformed into the original space to obtain an estimate of the center and a positive definite covariance matrix.

Another idea from (Gnanadesikan and Kettenring 1972) is related to the so-called forward search methods, which are closely related to the method proposed in this paper. These so called forward search methods are based on the concept of "growing a good subset of observations". By "good subset" one means a subset free or almost free of outliers. The idea is to start with a small subset of the data and then to add non-outlying observations until no more non-outliers are available.

The idea of a forward search algorithm was first suggested in (Wilks and Gnanadesikan 1964) and described in detail in (Gnanadesikan and Kettenring 1972). The articles of (Hadi 1992) and (Atkinson 1993) demonstrated the efficiency of such methods. In both articles the "good subset" grows one point at a time using Mahalanobis distances to rank the observations. Then research concentrated on developing faster and more sophisticated methods based on the same idea. The last two and most efficient were developed in (Billor, Hadi and Vellemann 2000) and (Kosinski 1999). These algorithms were compared in (Béguin 2002) and the BACON algorithm (Billor *et al.* 2000) turned out to be the most robust and fastest forward search method with complete multivariate normal data. In particular the breakdown point turned out to be high in practical applications. Also when comparing with other Mahalanobis type methods the performance of BACON on complete data is very good (Béguin and Hulliger 2003).

None of the above methods is designed to deal with incomplete data stemming from surveys, *i.e.*, with missing values and sampling weights. The first article to address the problem of multivariate outlier detection in incomplete data is (Little and Smith 1987). The authors propose Mahalanobis distances to detect outliers, with robust



estimations of center and scatter obtained by the ER algorithm. The ER algorithm replaces the maximum-likelihood estimator in the maximization step of the EM algorithm (Dempster, Laird and Rubin 1977) by a robust one-step  $M$ -estimator. However, the starting point of the ER algorithm is the classical non-robust mean and covariance and therefore the breakdown point of the ER algorithm is 0. In other words even one single outlier can carry the estimator beyond any limit. To correct the low breakdown point of that algorithm (Cheng and Victoria-Feser 2000) used an MCD algorithm for the maximization step of the EM-algorithm. However, the combination of the iterative procedures of MCD and EM makes the computation for large datasets too slow for practical applications. Moreover the introduction of sampling weights is not straightforward.

The TRC algorithm in (Béguin and Hulliger 2004) uses robust linear regression imputations by the best univariate predictor to cope with missing values. The Spearman rank correlations are expressed as functionals of the empirical distribution function of the sample to obtain estimates for the Spearman rank correlations in the population.

The BACON algorithm is based on the multivariate normal distribution and thus the EM algorithm for multivariate normal data was chosen to impute missing values within the BACON iterations. To take into account the sampling aspect, the estimates of the BACON algorithm have to be replaced by Horvitz-Thompson type estimators and a special version of the EM algorithm is developed where the expectations on the population level are estimated from the sample. Section 2 sets up the notation, recalls quickly the BACON algorithm and presents its adaptation to sampling weights. Section 3 introduces the Estimated-EM (EEM) algorithm and Section 1 discusses the adaptation of the Mahalanobis distance to missing values. Section 4 explains how BACON and EEM are merged in an efficient way to become the BACON-EEM algorithm. Section 5 shows the application of BACON-EEM to two datasets. The results are compared to the competitor methods, Transformed Rank Correlations, developed in (Béguin and Hulliger 2004), the ER-algorithm and a baseline algorithm which uses MCD after non-robust imputation based on the EM-algorithm.

2. The BACON algorithm

The BACON algorithm is presented in (Billor *et al.* 2000). Two versions are described: one for multivariate data in general and one for regression data. Only the first case will be considered here.

The data are stocked in a matrix  $X$  of  $n$  rows (observations  $x_1, \dots, x_n$ ) and  $p$  columns (variables  $x^1, \dots, x^p$ ). We assume that the bulk of the data is unimodal

and roughly elliptical symmetric. The coordinatewise mean (resp. covariance matrix) computed on  $X$  is denoted by  $m_X$  (resp.  $C_X$ ). The squared Mahalanobis distance of a point  $y$  based on  $m_X$  and  $C_X$  is  $MD_X^2(y) = (y - m_X)^\top C_X^{-1} (y - m_X)$ . If the mean and covariance are calculated only on a subset  $G$  of the data then we denote them  $m_G$  and  $C_G$  with corresponding Mahalanobis distance  $MD_G$ .

The first step of the algorithm is the choice of an initial subset  $G$  of “good data”. Two versions are proposed in the literature. The first version simply selects the  $cp$  points with smallest Mahalanobis distances  $MD_X(x_i)$ ,  $i \in \{1, \dots, n\}$ , with  $c$  being an integer chosen by the data analyst. It may be set to  $c=3$  by default. The second version selects the  $cp$  points with smallest Euclidean distances from the coordinatewise median, with  $c$  as before. The second version is more robust but it loses affine equivariance. Other starting points than the coordinatewise median might be considered like a spatial median. In this article we concentrate on the second version of the basic good subset. In both versions if  $C_G$  is singular then the basic subset is increased by adding observations with smallest distances until  $C_G$  has full rank. Then an iterative process starts.

Denote by  $\chi^2_{p,\beta}$  the  $1-\beta$  percentile of the  $\chi^2$  distribution with  $p$  degrees of freedom and by  $|G|$  the number of elements in the set  $G$ . The steps of the BACON algorithm are:

- 1. Compute the squared Mahalanobis distances  $MD_G^2(x_i)$  for  $i \in \{1, \dots, n\}$ ;
- 2. Define a subset  $G'$  including all points with  $MD_G^2(x_i) < c_{npr} \chi^2_{p,\alpha/n}$ , where  $c_{npr} = c_{np} + c_{hr}$  is a correction factor with  $c_{np} = 1 + (p+1)/(n-p) + 1/(n-h-p)$ ,  $c_{hr} = \max\{0, (h-r)/(h+r)\}$ ,  $h = \lceil (n+p+1)/2 \rceil$  and  $r = |G|$ .
- 3. If  $G' = G$  then stop, else set  $G$  to  $G'$  and go to Step 1.

Note that the correction factor  $c_{npr}$  is close to 1 for large  $n$ . The observations that are not contained in the final  $G$  are declared outliers. Alternatively a threshold for the Mahalanobis distance, above which observations are nominated outliers, may be chosen by inspecting the distribution of the Mahalanobis distance.

The computing effort required by the BACON algorithm depends on the configuration of the data. Compared with other algorithms it is small and in particular this effort grows slowly with increasing sample size (see also Section 5). This makes the BACON method particularly well suited for large datasets.

Note that the original selection criterion of Step 2 is designed for a multivariate normal distribution, which

implies that the squared Mahalanobis distances follow asymptotically a  $\chi^2$  distribution with  $p$  degrees of freedom. Suppose all points follow a multivariate normal distribution and that the Mahalanobis distance is computed using the sample mean and covariance matrix. The test  $MD_X^2(x_i) > \chi_{p,\alpha}^2$  declares about 100 $\alpha$  percent of the points as outliers. Instead of  $\alpha$  we often use  $\alpha/n$ . Using Bonferroni inequalities one can show that under normality the test with level  $\alpha/n$  will declare no outlier with probability larger than  $1 - \alpha$  (i.e.,  $P(MD_X^2(x_i) < \chi_{p,\alpha/n}^2, \forall i \in \{1, \dots, n\}) \geq 1 - \alpha$ ). The test with  $\alpha/n$  very rarely detects points that are not outliers but it also reduces its sensitivity to close outliers when  $n$  becomes large. One may also want to run the method with both types of the test level and compare the results.

## 2.1 Adaptation to sampling weights

For the sampling context we use the following notation. The data stem from a random sample  $s$  of the finite population  $U$  with  $N$  elements. The sample of size  $n$  is drawn with the sample design  $p(s)$  and the first order inclusion probabilities are denoted  $\pi_i = \sum_{s|i \in s} p(s)$ . The weights will be the inverse of the inclusion probabilities of the observations  $w_i = 1/\pi_i$  such that the Horvitz-Thompson estimator of the population total,  $\sum_{i \in U} x_i$ , is  $\sum_s w_i x_i = \sum_{i=1}^n w_i x_i$ . Furthermore it is assumed that  $\sum_s w_i \approx N$ . The mean  $m_X$  and the covariance matrix  $C_X$  may be estimated by the Hájek estimators

$$\hat{m}_X = \frac{\sum_s w_i x_i}{\sum_s w_i} \quad (1)$$

and

$$\hat{C}_X = \frac{\sum_s w_i (x_i - \hat{m}_X)(x_i - \hat{m}_X)^\top}{\sum_s w_i}.$$

The sample estimate of the median is defined as in (Béguin and Hulliger 2004): let  $x_u^k$  be the smallest value such that  $\sum_s w_i 1_{x \leq x_u^k}(x_i^k) \geq 0.5 \sum_s w_i$  and  $x_v^k$  the smallest value such that  $\sum_s w_i 1_{x \leq x_v^k}(x_i^k) > 0.5 \sum_s w_i$ , then the estimate is given by

$$\widehat{\text{med}}_X = (w_u x_u^k + w_v x_v^k) / (w_u + w_v). \quad (2)$$

To adapt the BACON algorithm to sampling the initial subset is selected using Hájek estimators  $\hat{m}_X$  and  $\hat{C}_X$  or the median  $\widehat{\text{med}}_X$ . For the iterative process, denote by  $s_G$  the selected “good observations” of the sample. These observations are representatives of a “virtual good subset”  $G$  of the whole population with estimated size  $\hat{r} = \sum_{s_G} w_i$ . The mean and covariance matrix of this subset are estimated by the Hájek estimators

$$\hat{m}_G = \frac{\sum_{s_G} w_i x_i}{\sum_{s_G} w_i} \quad (3)$$

and

$$\hat{C}_G = \frac{\sum_{s_G} w_i (x_i - \hat{m}_G)(x_i - \hat{m}_G)^\top}{\sum_{s_G} w_i}.$$

These estimates are used to compute the estimates of the Mahalanobis distances  $MD_G(x_i)$ ,  $x_i \in s$ . Finally the correction factor  $c_{Npr} = c_{Np} + c_{hr}$  of the selection criteria is computed using the estimates  $\hat{N} = \sum_{s_G} w_i$  and  $\hat{r} = \sum_{s_G} w_i$ . If  $N$  is known, its actual value is used.

If there are no missing values in the data the BACON algorithm can be used to estimate the population mean and covariance. The basic assumption for the BACON algorithm is still that the bulk of observations of the population has an elliptical distribution. We may use the BACON algorithm without weighting and compare the result with the weighted version. Different results indicate that the design-variables or a model used for non-response weighting are not well reflected in the model. We advocate the use of weights, in particular in routine applications, to give some protection against miss-specification of the model. In any case the estimand should be the mean and covariance of the bulk of the population.

Note that the Mahalanobis distance does not involve the sampling weights directly. The weight of a possible outlier influences the Mahalanobis distance only through the model, i.e., the mean and the covariance.

## 3. The EEM algorithm

Nonresponse issues are important in official statistics and many surveys cannot deliver a complete dataset. The problem of unit-nonresponse, i.e., completely missing observations, is usually dealt with by using appropriate weights and is not treated here. Item-nonresponse, i.e., observations with only partially available information, cannot be treated by discarding all incomplete observations because too much information is lost. The approach followed here will retain high efficiency under multivariate normal data. At each BACON iterative step the mean and covariance matrix of the good subset of observations will be computed using a modified version of the EM algorithm for multivariate normal data. The expectations computed in the E-step are replaced by sample estimates. The modified algorithm is therefore named the EEM (Estimated-Expectation/Maximization) algorithm. Note that this adaptation is presented here for multivariate normal data but the results can be generalized to other distributions of the regular exponential family.

This paragraph re-uses the description and notation of the EM algorithm given in (Schafer 2000). All details about EM not given here can be found within the first three chapters and in Section 5.3 of this book. The following abuse of notation is also used here:  $X$  will denote simultaneously a  $p$ -dimensional random variable and the  $N \times p$  matrix containing the realized values of the variable  $X$  of the population  $U$ . If a census were taken of the whole population to measure the variable  $X$  it would result in some observed and missing values  $X = X_o \cup X_m$ . The EM-algorithm assumes that the missingness mechanism is ignorable (Schafer 2000, section 2.2). Here we assume in addition that the missingness is independent from the sampling. The observations of the data can be modeled as independent, identically distributed (iid) draws from a multivariate normal probability distribution with density  $f(x, \theta)$ . Using the assumptions and the factorization  $P(X|\theta) = P(X_o|\theta)P(X_m|X_o, \theta)$  the complete-data log-likelihood can be written as  $l(\theta|X) = l(\theta|X_o) + \log(P(X_m|X_o, \theta)) + c$ , where  $l(\theta|X_o)$  is the observed-data log-likelihood and  $c$  is an arbitrary constant. The term  $P(X_m|X_o, \theta)$  captures the interdependence between  $X_m$  and  $\theta$  on which the EM-algorithm capitalizes. Because  $P(X_m|X_o, \theta)$  is unknown the average of  $l(\theta|X)$  over  $P(X_m|X_o, \theta^{(i)})$  is taken at each E-step, where  $\theta^{(i)}$  is a preliminary estimate of the unknown parameter. The next estimate  $\theta^{(i+1)}$  is found by maximizing the result of the expectation step (M-step). The sequence of E and M-steps is iterated until convergence. Conditions under which this sequence  $\theta^{(i)}$  converges to a stationary point of the observed-data likelihood are provided in (Dempster *et al.* 1977). In well-behaved problems this stationary point is a global maximum.

For a probability distribution of the regular exponential family the complete data log-likelihood may be written as

$$l(\theta|X) = \eta(\theta)^T \cdot T(X) + Ng(\theta) + c, \quad (4)$$

where  $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), \dots, \eta_k(\theta))^T$  is the canonical form of the parameter  $\theta$  and  $T(X) = (T_1(X), T_2(X), \dots, T_k(X))^T$  is the vector of complete-data sufficient statistics. Moreover, each of the sufficient statistics has an additive form  $T_j(X) = \sum_{i=1}^N h_j(x_i)$ , for some function  $h_j$ . Because  $l(\theta|X)$  is a linear function of the sufficient statistics, the E-step replaces  $T_j(X)$  by  $E(T_j(X)|X_o, \theta^{(i)})$ . In other words the E-step fills in the missing portions of the complete-data sufficient statistics. For a multivariate normal distribution  $X = (X^1, \dots, X^p)$  the sufficient statistics are composed of two types of elements: the sums  $\sum_{i=1}^N x_i^k$  and the sums of products  $\sum_{i=1}^N x_i^k x_i^l$ ,  $1 \leq k, l \leq p$ . The E-step reduces to computing the conditional expectations of these sums given the observed data  $X_o$  and the preliminary parameter  $\theta^{(i)}$ .

For a single summand  $i$  one can show (Schafer 2000, section 5.3) that these expectations depend only on the observed components of the same observation, *i.e.*, on  $x_i^{\text{obs}}$ . This leads to

$$\begin{aligned} E\left(\sum_{i=1}^N x_i^k \middle| X_o, \theta^{(i)}\right) &= \sum_{i=1}^N E(x_i^k | X_o, \theta^{(i)}) \\ &= \sum_{i=1}^N E(x_i^k | x_i^{\text{obs}}, \theta^{(i)}), 1 \leq k \leq p \end{aligned} \quad (5)$$

and the analogue form of the sum of products. Of course  $E(x_i^k | x_i^{\text{obs}}, \theta^{(i)}) = x_i^k$  if  $x_i^k \in x_i^{\text{obs}}$ . If  $x_i^k$  is missing, then this expectation is the fitted value of a regression of  $x^k$  given the parameter  $\theta^{(i)}$  on the variables which are observed for observation  $i$ . Thus the sufficient statistics are composed of population sums of observed values ( $T_o$ ) and sums of fitted values ( $T_m$ ).

In the situation where our data stem from a sample of a finite population we consider the finite population as a realization of a multivariate normal distribution and the sums (5) and sums of products have to be estimated from the sample. The form of (5) allows the use of simple Horvitz-Thompson estimators. The estimate of (5) is

$$T^{k0} = \sum_s w_i E(x_i^k | x_i^{\text{obs}}, \theta^{(i)}), 1 \leq k \leq p, \quad (6)$$

and  $E(\sum_{i=1}^N x_i^k x_i^l | X_o, \theta^{(i)})$  is estimated by

$$T^{kl} = \sum_s w_i E(x_i^k x_i^l | x_i^{\text{obs}}, \theta^{(i)}), 1 \leq k, l \leq p. \quad (7)$$

In short: We replace the population sums of  $T_o$  and  $T_m$  by their Horvitz-Thompson estimators  $\hat{T}_o^{k0}$  and  $\hat{T}_m^{kl}$ . We call the calculation of the  $T^{k0}$  and  $T^{kl}$  the estimated expectation step (EE-step). Plugging these estimators into (4) we obtain an estimator of the average population likelihood function.

For the M-step, the maximization of the estimate of the average population likelihood, the weighted normal equations have to be solved. The solution is found by a simple matrix operation using the sweep operator (Schafer 2000, section 5.3) applied to the symmetric  $(p+1) \times (p+1)$  matrix  $(T^{kl})_{0 \leq k, l \leq p}$  of the estimated expectations of the sufficient statistics (with  $T^{00}$  set to 1) divided by  $N$ , which is estimated by the sum of weights if unknown:

$$\theta^{(i+1)} = \text{SWP}[0] \left( \frac{(T^{kl})_{0 \leq k, l \leq p}}{\sum_s w_i} \right), \quad (8)$$

where  $\text{SWP}[0]$  is the sweep operator on the first line/column of the matrix.

The EEM algorithm iterates the EE and the M-step. Computationally the difference between the EE-step and the E-step of the original EM-algorithm comes down to using



weighted sums instead of un-weighted sums with weights that do not change over the iterations. We therefore expect that the convergence of the EEM-algorithm will remain similar to the EM-algorithm. For the BACON-EEM algorithm we only need a rough approximation to the solution in each BACON-step. Thus we use only a small number of iterations of the EEM algorithm.

### 3.1 Mahalanobis distance with missing values

The Mahalanobis distance is developed for complete observations and needs to be adapted to missing values. One option is to use the EEM estimate to impute the conditional mean for the missing values given the observed values and then calculate the Mahalanobis distance with imputed values. Under a MAR (Missing At Random) assumption there is a valid model based on the observed part of the data to impute the missing values. In the case of outlier detection we suppose that the imputation model may hold for the bulk of the data only and is estimated in a robust way. But then we may not expect that an outlier value is predicted by the model, except if already the observed part of an observations is outlying. Therefore there is no advantage to use imputation before outlier detection and we prefer to directly adapt the Mahalanobis distance to missingness. Two different versions of the Mahalanobis distance are possible in this situation.

We call the first version *marginal* Mahalanobis distance. It uses the Mahalanobis distance in the space of observed variables and scales it up with a factor  $p/q$ , where  $q = \sum_k r_{ik}$  is the number of non-missing variables and  $p$  is the total number of variables. More precisely, we assume an observation  $x$  is partitioned into  $x = (x_o^\top, x_m^\top)^\top$  (after possible rearrangement), where  $x_o$  denotes the observed part and  $x_m$  the unobserved part of the observation. Then the marginal Mahalanobis distance is

$$\text{MD}_{\text{marg}}^2 = \frac{p}{q} (x_o - m_o)^\top (S_{oo})^{-1} (x_o - m_o), \quad (9)$$

where  $S_{oo}$  is the part of the covariance matrix corresponding to  $x_o$ . This version is also used in (Little and Smith 1987).

The second version of Mahalanobis distance with missing values is obtained by reducing the contribution of the missing values to the Mahalanobis distance to zero. This amounts to replacing all missing values by their mean, i.e.,  $x_m = m_m$ . In other words we would impute a mean without consideration of the covariance matrix and the above arguments against outlier detection with imputed values apply here as well. Nevertheless we tested this second version of Mahalanobis distance. It yields erratic Mahalanobis distances (Béguin 2002) and (Béguin and Hulliger 2003) and we did not use it any further.

## 4. The BACON-EEM algorithm

Both algorithms, BACON and EEM, are computationally demanding. By merging them in a convenient way we gain performance. The “growing” structure of the BACON algorithm implies redundancies which may be used to avoid extra-computations in the EEM-algorithm at each step. The crucial point at each BACON step is that the estimations of the mean and the covariance matrix from the EEM-algorithm allow the exclusion of outlying points from the good subset and this does not need extremely precise estimates. Thus it is not necessary to iterate EEM to convergence each time the mean and covariance are needed. We use only 5 iterations by default. Furthermore we use the result of the last EEM-iteration of the last BACON-step as a starting value for EEM.

As much information from past iterations as possible should be reused. In fact the sufficient statistic  $T^G$  computed on some good subset  $G$  have an observed part of the sum  $T_o^G$  and a missing part of the sum  $T_m^G$ . The expectation computed by the E-step can therefore be written as

$$E(T^G | X_o^G, \theta) = T_o^G + E(T_m^G | X_o^G, \theta). \quad (10)$$

As the subsets  $G$  are usually growing,  $\hat{T}_o^G$  is not recomputed at each step of the BACON loop, but a global variable for  $\hat{T}_o^G$  is updated each time  $G$  changes (usually only adding points, sometimes removing a few).

At each iteration of the BACON-EEM algorithm, once the EEM algorithm has obtained the estimations of the center and the scatter of the good subset, marginal Mahalanobis distances for all observations are used in step 2 of the BACON algorithm.

Note the crucial point for the robustness of the algorithm: EEM is not robust, but at each BACON-step EEM is run only on points that have the smallest and therefore non-outlying marginal Mahalanobis distance in the preceding step. In other words the observation  $x$  will be used by EEM if and only if  $x_o$  is sufficiently small for the metric given by  $(S_{oo})^{-1}$  at the preceding step. Therefore if the first subset of good points is free or almost free of outliers, the imputation process in EEM will never create outlying values throughout the whole BACON-EEM algorithm. In other words, the non-robust EEM-algorithm is protected by the general forward search approach of the BACON algorithm in the same way as the non-robust mean and covariance of the original BACON algorithm is protected.

Summing up, the steps of the BACON-EEM algorithm are the following:

1. Calculate the weighted coordinate-wise median  $\widehat{\text{med}}(x)$  ignoring missing values in each variable separately. Determine the Euclidean distance from the median of

each observation omitting missing values but standardizing for the number of present values:  $a_i = \|x_i - \text{med}(x)\| \sqrt{p/q}$ . Select the  $m = cp$  observations with least  $a_i$  to constitute the initial subset  $G$ .

2. Compute a center  $\hat{m}_G$  and scatter  $\hat{C}_G$  using the EEM-algorithm and update the estimate of the sufficient statistic of the observed part  $\hat{T}_o^G$ .
3. Compute the squared marginal Mahalanobis distances  $MD_G^2(x_i)$  for  $i = 1, \dots, n$ . The new set  $G'$  contains the observations with  $MD_G^2(x_i) < c_{\hat{N}_p^G} \chi_{p,\alpha}^2$ .
4. If  $G' = G$  then stop, else set  $G$  to  $G'$  and go to step 2.

If instead of outlier detection the mean and covariance estimates of BACON-EEM are the main objectives the EEM-algorithm may be iterated further without changing  $G$ . In step 3 one may alternatively use  $\alpha/n$  instead of  $\alpha$  (see Section 2).

5. Applications

In this section we compare the BACON-EEM algorithm (BEM) with Transformed Rank Correlations (TRC) from (Béguin and Hulliger 2004) and the ER-algorithm from (Little and Smith 1987). As a further benchmark we use an imputation under the multivariate normal model with estimates of the mean and covariance by the EM algorithm. In other words we create a non-robust imputation. Then robust estimates of the multivariate location and the covariance matrix are obtained by the Minimum Covariance Determinant estimator computed on the imputed data and finally outliers are detected using the corresponding Mahalanobis distances. The benchmark method is called GIMCD for “Gauss Imputation followed by MCD

detection”. The algorithms are implemented in R (R Development Core Team 2006) with the help of the R-packages `norm` (Novo and Schafer 2002) and `MASS` (Venables and Ripley 2002).

5.1 Bushfire data

The reaction of the BACON-EEM to the introduction of missing values is illustrated with a real dataset of 38 observations and 5 variables. It was used by (Maronna and Zamar 2002) to locate bushfire scars. This well known example is also studied in (Maronna and Yohai 1995) and (Maronna and Zamar 2002). It allows a two dimensional plot (in variable 2 and 3) that reveals most of the outliers (see Figure 1). The data contains an outlying cluster of observations 33 to 38 a second outlier cluster of observations 7 to 11 and a few more isolated outliers, namely observations 12, 13, 31 and 32. We have added observation 31 to the list of potential outliers because it is indicated as a borderline case by MCD, BACON and also other methods studied in (Maronna and Zamar 2002). Missing values are created with a MCAR (Missing Completely At Random) mechanism. Two datasets are created with respectively 20 and 40% of missing items. The dataset with 40% of missing values have observations with up to 4 out of 5 missing values and therefore are a challenge for any method. As the size  $n$  of the dataset is small, BACON-EEM is run with the  $\chi_{p,\alpha/n}^2$  test. The results are given in Table 1. Observations 7 to 13 and 31 to 38 are individually shown as detected or not, while for the other 23 good points the number of observations declared as outliers is indicated. The limit above which a Mahalanobis distance indicates an outlier, was determined for each run by inspection of the quantile plot of the Mahalanobis distance.

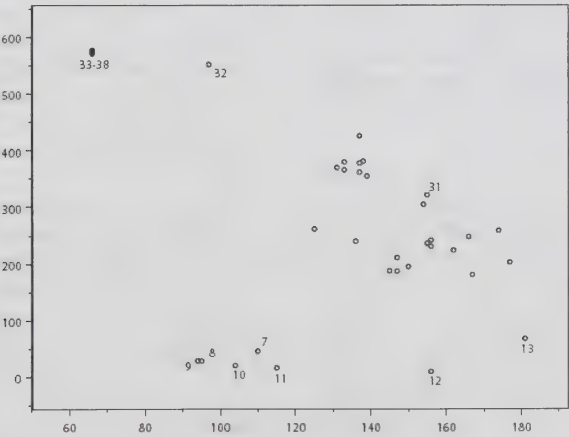


Figure 1 Bushfire Data

With complete observations none of the methods declared any good observations as outlier. The Mahalanobis distance with non-robust mean and covariance detects 3 outliers but misses the others. MCD and BACON-EEM end up with the same subset of data as good points and therefore give the same result, *i.e.*, exactly the same Mahalanobis distance for all observations. Both do not declare observations 12 and 13 as outliers but all the others. ER does detect the group of outliers 7,...,12 but none of the other outliers. TRC detects the group 32,...,38 and two more outliers.

With 20% missing values BEM and TRC declare one good data point as outlier. TRC detects 14 of the 15 potential outliers. ER misses all outliers except observation 7. GIMCD detects the same 13 outliers as without missing values. Note however, that there is some variability in the results for GIMCD due to the random imputation. BEM detects the same outliers as with complete data except observation 11.

Table 1 Outliers detected for 3 missingness rates

(1-q)%	Method	7-11	12,13,21	32-38	n. good
0	MD	11100	000	0000000	0
0	MCD	11111	001	1111111	0
0	ER	11111	100	0000000	0
0	BEM	11111	001	1111111	0
0	TRC	01100	000	1111111	0
20	GIMCD	11111	001	1111111	0
20	ER	10000	000	0000000	0
20	BEM	11111	000	1111111	1
20	TRC	01111	111	1111111	1
40	GIMCD(1)	11100	000	1111111	0
40	GIMCD(2)	11100	000	0100000	5
40	ER	10000	010	0000000	2
40	BEM	11111	000	1111111	1
40	TRC	11111	010	1111111	1

MD: Classical Mahalanobis distance, MCD: Minimum Covariance Determinant, GIMCD: Non-robust imputation under Gaussian model followed by MCD (GIMCD(1) and GIMCD(2) are two realisations of the GIMCD-algorithm), ER: Expectation-Maximization with one M-step at maximization, BEM: BACON-EEM, TRC: Transformed Rank Correlations. The first column indicates the proportion of missing values, the last column gives the number of other points (non-outliers) declared outliers, the intermediate columns are detection indicators for the observations in the first row.

With 40% missing values ER nominates observation 7, 13 and two good observations as outliers. Since the imputation is random the result of GIMCD has some variability. Two realizations, GIMCD(1) and GIMCD(2) are reported in Table 1. In a good case GIMCD detects 10 of the 15 outliers and does not declare any good observation as outlier. In a bad case GIMCD detects only 4 outliers but declares 5 good observations as outliers. BEM detects 12 of the outliers and declares one good observation as outlier. TRC detects 13 outliers and declares one good observation as outlier.

## 5.2 MU281 data

The MU284 data set from (Särndal, Swensson and Wretman 1992) contains data about Swedish municipalities. We use the variables *population in 1975* and *population in 1985* (pop75 and pop85), *revenue from municipal taxes 1985* (RMT85), *number of municipal employees 1984* (ME84) and *real estate value 1984* (REV84). The largest three cities according to pop75 are discarded because they are huge outliers and would be treated separately in practice. The remaining municipalities are supposed to be a stratified sample of a larger population. Strata are defined according to  $0 < \text{pop75} < 20$ ,  $20 \leq \text{pop75} < 100$ ,  $100 \leq \text{pop75}$ . Table 2 shows the assumed population sizes and the corresponding weights. This sample design reflects a typical stratification for establishment surveys with a take-all stratum of the largest establishments, where in the end 8 of 10 establishments answer the survey.

Table 2 MU281 population and sample sizes

	stratum		
	1	2	3
pop75	0-19	20-99	100+
N	1,600	250	10
n	171	102	8
w	9.36	2.45	1.25

The three variables RMT85, ME84 and REV84 are divided by pop85 to obtain figures per capita. The per capita variables are denoted by lower case names (rmt85, me84 and rev84). Figure 2 shows the distribution of these 3 variables plus the auxiliary variable pop75. The per capita figures are roughly elliptically distributed. There is a linear relationship between rmt85 and me84 and a slightly non-linear relationship of these variables with pop75. There is no apparent relationship between rev84 and rmt85 and between rev84 and me84 but there is clearly more variability in rev84 for low pop75. The distributions of variable pop75 and rev84 are skew. There is a large outlier in rmt85 and me84 and at least two in rev84.

We include pop75 in all our calculations. In practice one would include the auxiliary variable which defines the sample design in a model. Note that pop75 has no missing values.

The qq-plot of Mahalanobis distances based on MCD shows only the two clear outliers in rev84. The large outlier jointly in rmt85 and me84 has 25th largest Mahalanobis distance. We call these largest 25 observations the unweighted basic outliers. In the original MU284 dataset these unweighted basic outliers have LABEL 3, 4, 29, 31, 46, 47, 56, 79, 83, 117, 126, 131, 140, 158, 199, 211, 222, 246, 248, 252, 254, 260, 262, 272, and 273. With classical non-robust Mahalanobis distances only 12 of the basic outliers are nominated outliers, *i.e.*, are among the 25 observations with



largest (classical) Mahalanobis distance. Robust methods are necessary to detect the outliers in the MU281 data. To allow a comparison between the methods we fix the number of observations which are to be considered outliers to 25 for the moment. Thus we consider for each method the 25 observations with largest Mahalanobis distance as the outliers. Note that this may not be the threshold one would choose after inspection of the qq-plot of the Mahalanobis distances.

Table 3 shows the number of basic outliers detected by the methods run on the complete dataset. MCD detects its own 25 outliers, of course. The ER, BEM and TRC algorithm detect 25 or 24 of these outliers if no weights are applied but only 11 or 15 if weighted. A suffix “w” behind the acronym of the method indicates that the sampling weights were used. Since the small municipalities have more weight the estimates are attracted towards them and other outliers will come up among the 25 largest Mahalanobis distances for ERw, BEMw and TRCw (see also Table 4). Closer inspection shows that many of the unweighted outliers are located in the tail of rev84 while most of the weighted outliers are located in the tail of pop75. The weighted methods coincide on 20 observations as outliers. We will call them weighted basic outliers. The weighted basic outliers have original MU284-LABEL 16, 28, 36, 45, 46, 55, 97, 113, 115, 121, 155, 185, 196, 208, 233, 241, 245, 265, 267, and 270. Only 10 of these observations are also among the unweighted basic outliers.

Table 4 shows the number of weighted and unweighted basic outliers in the strata. There are 12 unweighted but only 2 weighted basic outliers in stratum 1. Thus the weights have a clear influence on outlier detection. The influence is on the model primarily which is attracted towards the small observations with larger weights. Of course this can be seen as a sort of masking of outliers but in the context of modeling a better explanation is that the model is not completely adequate over all the strata and the weighted model fits the population better than the unweighted.

The second row of Table 3 gives the computation time for the algorithms. The ER algorithm is much slower than its competitors. This may be due to an inefficient implementation, however. The fastest algorithm is BEM, followed by TRC and, at some distance MCD. TRC may become slow, however, when the missingness rate is high.

Table 3 Complete MU281, number of detected unweighted basic outliers

Method	MCD	ER	ER <sub>w</sub>	BEM	BEM <sub>w</sub>	TRC	TRC <sub>w</sub>
Number detected	25	25	11	24	15	24	15
Computation time	0.81	3.17	2.52	0.07	0.04	0.14	0.14

Suffix w indicates that the algorithm is run with sampling weights.

Table 4 Number of basic outliers per stratum

stratum	1	2	3	Total
unweighted	12	5	8	25
weighted	2	10	8	20

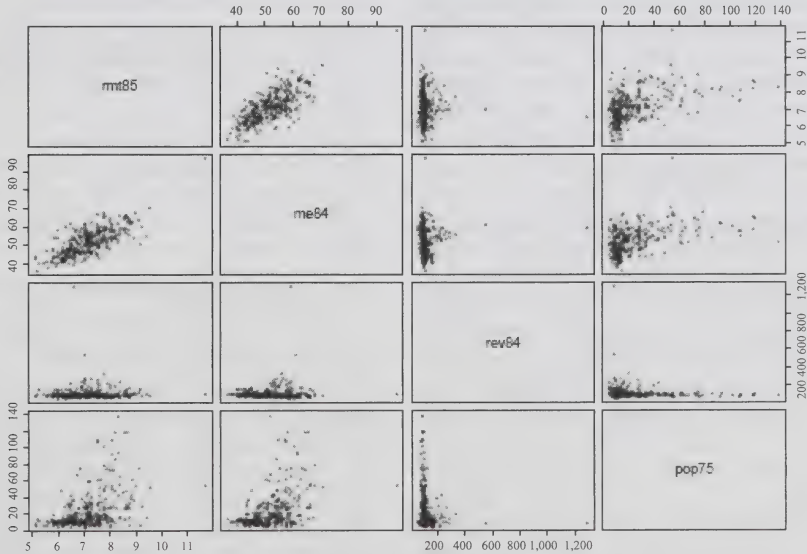


Figure 2 MU281 per capita figures and pop75

5.2.1 Missing values

We now introduce missingness into the variables rmt85, me84 and rev84 according to a mild MAR mechanism. Missingness patterns are assigned to the observations according to the modulo of their label in the original MU284 data. The MAR mechanism is reflected in higher missingness rates for stratum 1 and 2 (see Appendix). The response patterns and missingness rates per stratum are shown in Table 5. For example rev84 is the only missing value in 15 observations of stratum 1 and 2 of stratum 2. Overall 187 observations remain complete and the proportion of observations with missing values (missingness rate) is 33%.

Table 5 Frequency of response patterns per stratum for rmt85, me84, rev84

Response indicator			stratum		
rmt85	me84	rev84	1	2	3
0	1	1	11	4	0
1	0	1	13	2	0
1	1	0	15	2	0
1	0	0	13	2	1
0	1	0	14	2	0
0	0	1	11	4	0
1	1	1	94	86	7
missingness rate			0.450	0.157	0.125

Among the 35 weighted or unweighted basic outliers there are 17 observations with missing values. Table 6 shows how many of the basic outliers have been detected after the introduction of missingness. The 20 weighted basic outliers are detected well by the weighted algorithms ERw, BEMw and TRCw. GIMCD detects 4 of the weighted basic outliers and 14 of the unweighted basic outliers. Thus the missingness affects the capability of the MCD algorithm which was actually used to define the unweighted basic outliers. One word of caution: Several runs of random Gaussian imputation have been made and there is some variability in the results of GIMCD. However, also with a favorable imputation outcome GIMCD did not beat ER, BEM or TRC in detecting the unweighted basic outliers. The weighted versions ERw, BEMw and TRCw detect the weighted basic outliers well. The number of complete observations among the outliers nominated by the different methods is indicated in the last row of Table 6. All methods nominate as outliers also observations with missing values. Since the missingness rate is larger in the stratum of small observations and the weighted versions of the methods nominate less outliers in this stratum, the number of complete outliers is usually larger for the weighted algorithms. Overall the introduction of missingness has not altered the capabilities of ER, BEM and TRC by much, while GIMCD is moderately affected.

Table 6 MU281 data set with missing values, number of detected basic outliers

Method	GIMD	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Weighted	14	4	10	20	9	19	12	17
Unweighted	12	14	23	11	22	15	22	17
Complete	16	8	17	20	16	18	19	18

GIMD: Non-robust imputation under a Gaussian model followed by classical Mahalanobis distance.

5.2.2 Additional outliers

In addition to the outliers in the original data we now introduce new outliers. The observations which should become additional outliers are determined by the modulo of the original LABEL. If (LABEL mod 8=1 and pop75 ≥ 10) or (LABEL mod 16=1 and pop75 < 10) then the observation is an additional outlier. Thus the rate of outlyingness is larger for large municipalities. But the outlyingness is not influenced by the values of the other variables. We may say that the outlyingness is at random. Note that we could have taken a random sample instead of the above systematic sample. We preferred the systematic sample to simplify the replication of the results and to avoid additional randomness.

Two of the weighted and one of the unweighted basic outliers happen to be also additional outliers. We continue to treat them as basic outliers. Taking this into account there are 32 additional outliers in the sample. Together with the 25 unweighted or the 20 weighted basic outliers defined above there are 57 or 52 outliers to detect (20.3% or 18.5% outliers). From now on the threshold for the Mahalanobis distances is set at the 57<sup>th</sup> largest distance to simplify the comparison of the methods.

The values of the additional outliers are created as follows: rmt = 0.2 \* rmt85 + 8, me = 0.1 \* me84 + 50, rev = 0.4 \* rev84 + 300. Note that we omit the suffix indicating the year for the contaminated variables. The dependence on the old values is negligible. It is only used to avoid an explicit model for the error around the point (rmt, me, rev) = (8, 50, 300). This is the type of contamination that is difficult to detect for robust covariance estimators (Rocke and Woodruff 1996): concentrated and close to the point cloud of good observations.

Figure 3 shows the three variables with contamination and the location of the additional outliers.

Table 7 shows the number of detected outliers. GIMCD detects 31 of the 32 additional outliers, while BEM, BEMw, TRC and TRCw detect many of them but not all. ER and ERw detect less of the additional outliers. The weighted basic outliers are all detected by ER and ERw, BEMw and TRCw. The unweighted versions of BEM and TRC detect less of the weighted basic outliers and GIMCD detects only 4 of the weighted basic outliers. BEM and TRC whether weighted or not detect the unweighted basic outliers best.

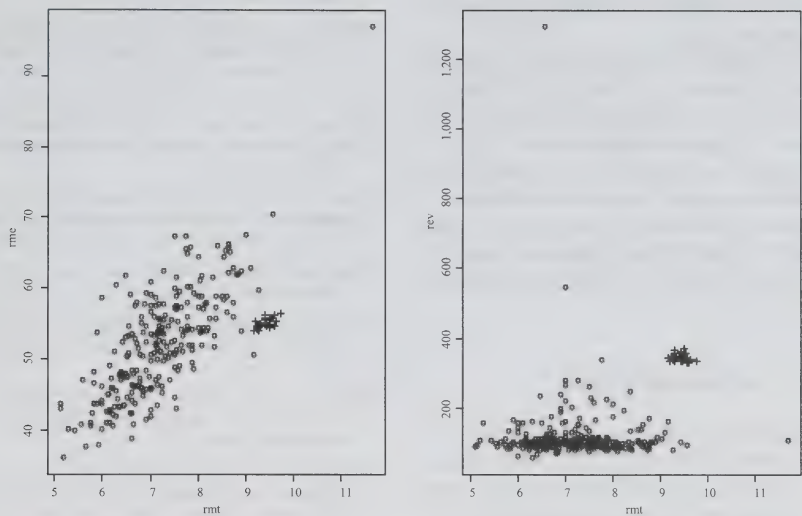


Figure 3 Variables me and rev vs. rmt. Additional outliers are marked with a cross

The last row in Table 7 shows that its non-robust imputation leads GIMCD to nominate more outliers with missing values than the other methods which robustify their imputations already before the detection phase.

Table 7 MU281 with missingness and moderate additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	32	31	19	6	27	27	28	27
Weighted	20	4	20	20	11	20	12	20
basic								
Unweighted	25	13	15	12	23	19	23	18
basic								
Complete		24	34	43	38	40	40	40

n. out: number of outliers, n. complete: number of complete observations among detected outliers.

In order to check the breakdown of the methods when a high number of additional outliers contaminates the data we set outliers if (LABEL mod 2 = 0 and pop75 ≥ 20) or (LABEL mod 3 = 1 and pop75 < 20). Excluding observations which are already basic outliers there are 98 additional outliers. Thus together with the 25 unweighted basic outliers we obtain 43.8% outliers. The threshold for the methods is therefore set at the 123<sup>rd</sup> largest Mahalanobis distance. Table 8 shows that, due to this large threshold, all weighted basic outliers are detected by all methods. The methods GIMCD, ER and ERw cannot cope with the high rate of outliers. BEM detects most of the outliers with BEMw and TRC only slightly behind. TRCw detects

somewhat less of the unweighted basic outliers and of the additional outliers.

Table 8 MU281 with missingness and high additional contamination

Method	n. out	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Additional	98	20	19	37	85	85	85	80
Basic	20	20	20	20	20	20	20	20
weighted								
Basic	25	21	19	17	23	18	18	13
unweighted								

6. Conclusions

The EM-algorithm for multivariate normal data can be adapted to a sampling context. The BACON algorithm protects the non-robust EEM-algorithm from outliers when the latter is applied within an iteration of the BACON algorithm. The ER-algorithm uses robustification within the EM-algorithm. The applications showed that this may not yield enough robustification. A possible reason, however, may also be the non-robust starting point of the M-step in the ER-algorithm.

GIMCD, a non-robust EM-algorithm followed by an imputation and detection with MCD covariance worked remarkably well for moderate missingness and contamination. Its variability with high missingness rate is a disadvantage. More stable solutions which also can take into account the sampling design should be explored. The



BACON-EEM algorithm showed very good detection capabilities in particular when the missingness rate and the contamination rate are high.

In spite of its simplicity the TRC algorithm is a good method in many circumstances. Its main problem seems to be the *ad-hoc* imputation with only one covariable, which can be a problem with high missingness rates.

In order to find a good model for the population it is important to use the sampling weights. Nevertheless, it is advisable to use also a non-weighted version and to check the differences. It is possible that outliers are masked by large sampling weights because they may then dominate the model estimate.

### Acknowledgement

The EUREDIT research project was part of the Information Society Technology Program (IST) of Framework Program 5 of the European Union. The Swiss participation in EUREDIT was supported by the Swiss Federal Office for Education and Science. A large part of this research was carried out while both authors worked at the Swiss Federal Statistical Office. The authors wish to thank the referees and editors for their valuable remarks.

### Appendix

#### Missingness in MU281

The default response pattern is 111, indicating that the three variables rmt85, me84 and rev84 are all present. A 0 in the string indicates a missing value for the corresponding variable. First, for all strata the response pattern is changed according to the following scheme with parameters  $(a, b, c) = (1, 2, 3)$ :

$$\text{response pattern} = \begin{cases} 011, & \text{if LABEL mod } 20 = a; \\ 101, & \text{if LABEL mod } 20 = b; \\ 110, & \text{if LABEL mod } 20 = c; \\ 100, & \text{if LABEL mod } 30 = a; \\ 010, & \text{if LABEL mod } 30 = b; \\ 001, & \text{if LABEL mod } 30 = c. \end{cases}$$

Additionally, the above scheme with parameters  $(a, b, c) = (5, 6, 7)$  is applied for stratum 1 again.

### References

- Atkinson, A. (1993). Stalactite plots and robust estimation for the detection of multivariate outliers. In *Data Analysis and Robustness*. (Eds., S. Morgenthaler, E. Ronchetti and W. Stahel), Birkhäuser.
- Béguin, C. (2002). Outlier detection in multivariate data. Master's thesis, Université de Neuchâtel.
- Béguin, C., and Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT.
- Béguin, C., and Hulliger, B. (2004). Multivariate outlier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations. *Journal of the Royal Statistical Society, A* 167(Part 2), 275-294.
- Billor, N., Hadi, A.S. and Vellemann, P.F. (2000). BACON: Blocked Adaptive Computationally-efficient Outlier Nominators. *Computational Statistics and Data Analysis*, 34(3), 279-298.
- Campbell, N. (1989). Bushfire mapping using noaa avhrr data. Technical report, CSIRO.
- Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81(396), 1063-1069.
- Cheng, T.-C., and Victoria-Feser, M.-P. (2000). Robust correlation estimation with missing data. Technical Report 2000.05, Université de Genève.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society, Series B, Methodological*, 39, 1-22.
- Donoho, D. (1982). *Breakdown Properties of Multivariate Location Estimators*. Ph.d. qualifying paper, Department of Statistics, Harvard University.
- EUREDIT (2003). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1 and 2. EUREDIT consortium. <http://www.cs.york.ac.uk/euredit/results/results.html>.
- Gnanadesikan, R., and Kettenring, J.R. (1972, March). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28, 81-124.
- Hadi, A.S. (1992). Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, B*, 54(3), 761-771.
- Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons, Inc.
- Kosinski, A.S. (1999). A procedure for the detection of multivariate outliers. *Computational Statistics & Data Analysis*, 29, 145-161.
- Little, R., and Smith, P. (1987). Editing and imputation for quantitative survey data. *Journal of the American Statistical Association*, 82, 58-68.
- Little, R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
- Liu, R.Y., Parelius, J.M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27(3), 783-858.
- Maronna, R., and Zamar, R. (2002). Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, 44(4), 307-317.
- Maronna, R.A. (1976). Robust *M*-estimators of multivariate location and scatter. *The Annals of Statistics*, 4, 51-67.
- Maronna, R.A., and Yohai, V.J. (1995). The behaviour of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90(429), 330-341.
- Novo, A.A., and Schafer, J.L. (2002). *norm: Analysis of multivariate normal datasets with missing values*. R package version 1.0-9.

- R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.
- Rocke, D., and Woodruff, D. (1993). Computation of robust estimates of multivariate location and shape. *Statistica Neerlandica*, 47, 27-42.
- Rocke, D., and Woodruff, D. (1996). Identification of outlier in multivariate data. *Journal of the American Statistical Association*, 91(435), 1047-1061.
- Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, Volume B, 283-297. Elsevier.
- Rousseeuw, P.J., and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*. New York: John Wiley & Sons, Inc.
- Schafer, J. (2000). *Analysis of Incomplete Multivariate Data*, Volume 72 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.
- Särdal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer.
- Stahel, W. (1981). *Robuste Schätzungen: infinitesimale optimalität und Schätzungen von Kovarianzmatrizen*. Ph.D. Thesis, Swiss Federal Institute of Technology.
- Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S* (Fourth Ed.). New York: Springer. ISBN 0-387-95457-0.
- Wilks, S.S., and Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics*, 35, 623-631.





# Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias

Annette Jäckle and Peter Lynn<sup>1</sup>

## Abstract

Respondent incentives are increasingly used as a measure of combating falling response rates and resulting risks of nonresponse bias. Nonresponse in panel surveys is particularly problematic, since even low wave-on-wave nonresponse rates can lead to substantial cumulative losses; if nonresponse is differential, this may lead to increasing bias across waves. Although the effects of incentives have been studied extensively in cross-sectional contexts, little is known about cumulative effects across waves of a panel. We provide new evidence about the effects of continued incentive payments on attrition, bias and item nonresponse, using data from a large scale, multi-wave, mixed mode incentive experiment on a UK government panel survey of young people. In this study, incentives significantly reduced attrition, far outweighing negative effects on item response rates in terms of the amount of information collected by the survey per issued case. Incentives had proportionate effects on retention rates across a range of respondent characteristics and as a result did not reduce attrition bias in terms of those characteristics. The effects of incentives on retention rates were larger for unconditional than conditional incentives and larger in postal than telephone mode. Across waves, the effects on attrition decreased somewhat, although the effects on item nonresponse and the lack of effect on bias remained constant. The effects of incentives at later waves appeared to be independent of incentive treatments and mode of data collection at earlier waves.

**Key Words:** Attrition; Item nonresponse; Mail survey; CATI; Unconditional incentive; Conditional incentive.

## 1. Introduction

Surveys around the world are facing declining response rates and, with this, increasing risks of nonresponse bias if nonrespondents' characteristics systematically differ from respondents' characteristics. For panel surveys this is particularly problematic, since even low nonresponse rates at each wave can lead to large cumulative losses. If nonresponse is differential, bias could increase with the duration of the panel. In order to boost participation rates, survey organisations increasingly offer respondent incentives. This paper provides new evidence on the cumulative effects of incentives on attrition, attrition bias and item nonresponse, using data from a large scale, multi-wave, mixed mode incentive experiment on a UK government panel survey of young people.

The effects of incentives have been studied in many settings: monetary incentives increase response more than gifts or lotteries (Church 1993; Singer, Hoewyk, Gebler, Raghunathan and McGonagle 1999); unconditional incentives (*i.e.*, those incentives that are given at the time of the survey request) increase response more than conditional incentives (those that are promised in return for participation) (Church 1993; Goyder 1994; Hopkins and Gullickson 1992; Singer *et al.* 1999); response rates increase with the value of the incentive (Armstrong 1975; Church 1993; Fox, Crask and Kim 1988; Hopkins and Gullickson 1992; Rodgers 2002; Yu and Cooper 1983); incentives have larger effects in studies with low response rates and larger effects

in postal than interviewer administered surveys (Singer *et al.* 1999). Most evidence of differences between modes in the effect of incentives, however, stems from comparisons of separate studies and fails to control for differences in other measures affecting response. As a result, differences in the effects of incentives are not necessarily genuine mode effects. The study by Ryu, Couper and Marans (2006) is an exception. The authors compared the effects of monetary incentives and gifts in a mixed mode postal and face-to-face survey. Their study did not, however, include a no-incentive condition and so did not allow an evaluation of the magnitude of incentive effects across modes. We compared the effects of incentives in a mixed postal and computer-assisted telephone interviewing (CATI) survey and, in postal mode, also examined the effects of conditional and unconditional incentives.

Research on the effects of incentives has focused on response rates and little is known about the effects on bias, the ultimate reason for concern about low response. Incentive studies are mostly limited to studying effects on bias in sample composition and some studies have found that incentives disproportionately increase participation of respondents typically under-represented, for example those with low education (Singer, Van Hoewyk and Maher 2000), poor (James 1997), black or poor (Mack, Huggins, Keathley and Sundukchi 1998), of black or Indian minority ethnic groups, living in larger households or households with dependent children, aged 0-20, or single (Stratford, Simmonds and Nicolaas 2003). Biases in sample composition are

1. Annette Jäckle, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. E-mail: aejack@essex.ac.uk; Peter Lynn, Institute for Social and Economic Research, University of Essex, Colchester CO4 3SQ, UK. E-mail: plynn@essex.ac.uk.

however not necessarily correlated with biases in important survey estimates, especially since typically only a handful of compositional characteristics are studied. Ultimately, studies of nonresponse bias are limited by the lack of information about nonrespondents, a limitation which can be overcome to some extent by panel studies, where information about nonrespondents is available from waves prior to the dropout. We estimate the extent of bias due to attrition in terms of sample composition and survey variables. We then assess the effectiveness of incentives at reducing bias, exploiting the information on attriters available from the panel.

Additionally, little is known about the effects of incentives over waves of a panel survey, whether the same treatment is administered repeatedly or whether the treatment changes between waves. In a review of the use of incentives in longitudinal studies, Laurie and Lynn (in press) concluded that, given the cost implications of changing incentive conditions, there was surprisingly little evidence about the longer term effects of such changes to guide survey practitioners. Quoting an internal memorandum of the US Census Bureau, Ward, Boggess, Selvavel and McMahon (2001) wrote (see page 2) that a "review of the well-known longitudinal studies (Downs 1999) found that all non-Census Bureau studies used a monetary incentive during each wave, but there had been no scientific tests to determine the effectiveness of the incentives". If attrition leads to dropout of the least co-operative, the sample might increasingly be composed of committed respondents who are less responsive to incentives, because they are sufficiently motivated to participate even without the incentive (Laurie and Lynn in press). In this case, incentives may have decreasing marginal effects on response rates over the life of the panel. By the same token, incentives may have increasing effects on attrition bias, if they have disproportionate effect on sample members who would otherwise be more likely to drop out. Although some incentive studies have been carried out in the context of panels, they mostly only covered one wave, or examined the effect of changes in incentive treatment from one wave to the next. Martin, Abreu and Winters (2001) and Ward *et al.* (2001), for example, studied the effects of incentives on conversion rates of previous wave nonrespondents; Rodgers (2002) and Laurie (2007) examined the effects of changes in incentive values in a panel. The only studies which examined the effects of incentives over more than two waves appear to be those by James (1997), Mack, Huggins, Keathley and Sundukchi (1998) and Laurie and Lynn (in press), who reported that the positive effect of an incentive paid early in a panel persisted for several waves even without repeated incentive payment. These studies, however, only examined the effect of an incentive paid in a

single wave and did not examine the cumulative effects of incentives offered over successive waves. We examine the cumulative effects of continued incentive payments across three waves spanning a time frame of three years, as well as the effects of changes from telephone to postal mode and from conditional to unconditional incentive treatment.

Finally, there is conflicting evidence in the literature about the effects of incentives on data quality. Although concern is frequently voiced that incentives may lead to lower data quality, by marginally increasing the motivation of respondents who would otherwise have dropped out of the study and are not sufficiently able or motivated to respond diligently, existing studies have either found that incentives lead to improved respondent effort and less item nonresponse (James and Bolstein 1990; Mack *et al.* 1998; Singer *et al.* 2000), or have found no relationship (Berk, Mathiowetz, Ward and White 1987; Davern, Rockwood, Sherrod and Campbell 2003; Goyder 1994; Shettle and Mooney 1999; Singer *et al.* 1999; Teisl, Roe and Vayda 2005; Tzamourani and Lynn 1999; Willimack, Schuman, Pennell and Lepkowski 1995). Item nonresponse is potentially critical, because analysts typically only use cases with complete data. This leads to losses in efficiency due to reductions in sample sizes and, similar to unit nonresponse, can lead to biased estimates and invalid inference if item nonrespondents are not a random subset of the sample (Mason, Lesser and Traugott 2002). Problems of item nonresponse increase for multivariate analysis, if the patterns of missingness vary across items, and for analysis of change, which in addition depends on complete information at different points in time. Since incentives may affect both unit and item nonresponse, it is then not clear what their net effect may be on repeated measures derived from a panel study. We examined the effect of incentives on item nonresponse rates and calculated their net effect on attrition and item response.

## 2. Hypotheses tested

The outcomes measured for this analysis were the attrition rate, item nonresponse rate and attrition bias. Attrition was an absorbing state, since the survey did not re-issue nonrespondents at later waves. Item nonresponse was measured as the number of non-filtered items missing, either due to refusals or 'don't know' answers. (Non-filtered items are those which apply to all sample members: items for which eligibility is determined by the response to an earlier question are excluded from our measure of item non-response.) Attrition bias was measured in terms of socio-demographic characteristics and wave 1 survey measures. These three outcome measures were used to test the following:



*H1: Effects of incentives on attrition, item nonresponse and attrition bias.*

In previous studies incentives have generally increased response rates, be it because norms of social exchange oblige the respondent to return a “favour” (norm of reciprocity, Gouldner 1960) or because the incentive substitutes for a lack of motivation to participate for other reasons, such as civic duty or topic interest (leverage-salience theory, Groves, Singer and Corning 2000). Incentives may in addition motivate respondents to provide better quality responses, reducing item nonresponse. At the same time incentives may change the sample composition to include more respondents who are not diligent about answering the survey questions, and as a result increase item nonresponse. Finally, incentives may have differential effects on attrition across sample members. Those with a high propensity to participate in the survey without the incentive may be less likely to be affected by incentives, while those more likely to drop out of the survey may be more susceptible. As a result, incentives may reduce attrition bias.

*Null hypothesis H1: Incentives have no effect on attrition, item nonresponse or attrition bias.*

*H2: Effects of incentives across waves.*

The effect of incentives in increasing unit and item response rates may weaken across waves, if attrition leads to dropout of the least motivated sample members and the remaining members are sufficiently motivated to participate for other reasons and hence less susceptible to incentives (Laurie and Lynn in press). However, the extent to which incentives reduce non-response bias could increase over waves, if incentives disproportionately retain those in the sample who are most likely to otherwise drop out.

*Null hypothesis H2: The effects of incentives do not change across waves.*

*H3: Effects of unconditional and conditional incentives in a panel context.*

Previous studies, carried out on cross-sectional surveys, suggest that unconditional incentives have larger effects on unit nonresponse, possibly because the prepayment signals that the survey organisation trusts the sample member will participate, reinforcing the norm of reciprocity. Whether the different incentive conditions have different effects on item nonresponse is not clear.

*Null hypothesis H3: Unconditional and conditional incentives have similar effects in a panel context.*

*H4: Effects of incentives in postal and telephone mode.*

Comparisons of previous studies suggest that incentives have a larger effect in postal mode, possibly because in telephone mode the interviewer already functions as an external motivator to increase both unit and item response (Singer *et al.* 1999) and the scope for additional improvements is smaller. The same may not necessarily be true in a panel context where the effect of mode on response may be mediated by the respondent’s experience of previous waves.

*Null hypothesis H4: Incentives have similar effects in postal and telephone mode.*

*H5: Effects of changes over waves in mode or incentive treatment.*

Compared to sample members allocated to the same mode and treatment across waves, those who were allocated to different treatments or different modes may differ in their experiences of previous survey waves and their expectations about future waves. As a result, the effect of incentives may not only be conditional on mode at the current wave, but may be influenced by the incentive treatment and mode in previous waves.

*Null hypothesis H5: Changes in mode or incentive treatment over waves do not have lasting effects.*

*H6: Effects of incentives across ability levels.*

Sample members with low education levels are typically more likely to drop out of surveys. If incentives reduce attrition bias, they should therefore disproportionately reduce attrition among lower achievers. Low ability respondents may at the same time be more likely to provide incomplete responses, if they find the task of completing the postal questionnaire more difficult. Therefore, incentives may increase mean levels of item nonresponse.

*Null hypothesis H6: Incentives have similar effects across ability levels.*

**3. Study design**

The Youth Cohort Study of England and Wales (YCS) investigates transitions from compulsory education to further or higher education or the labour market and typically samples cohorts of 16 to 17 year-olds every two years, who are surveyed on several occasions at annual intervals. The incentives experiment was embedded in waves 2, 3 and 4 of cohort 10. The survey is managed and funded by the Department for Children, Schools and Families, who jointly designed the incentive experiment



with the National Centre for Social Research, the survey contractors for waves 2 and 3 of YCS cohort 10.

3.1 The survey

The population studied in the YCS cohort 10 consisted of pupils in England and Wales who had reached minimum school leaving age of 16 in the 1998/1999 school year (Russell and Phelps 2001), that is, a one year age cohort of pupils born between 1-9-1982 and 31-8-1983. A 10% random sample was drawn from the registers of schools (excluding special schools and schools with fewer than 20 pupils of that age) in 1999, by asking schools to provide the names and addresses of pupils born on the 5<sup>th</sup>, 15<sup>th</sup> and 25<sup>th</sup> of every month. From the resulting file of 31,424 names and addresses a systematic random sample of 25,000 pupils was drawn. The first wave of the survey took place a year later in spring 2000, the second at the end of 2000, the third in spring 2002 and the fourth in spring 2003. Nonrespondents were not issued in subsequent waves and, as a result, attrition was monotonic.

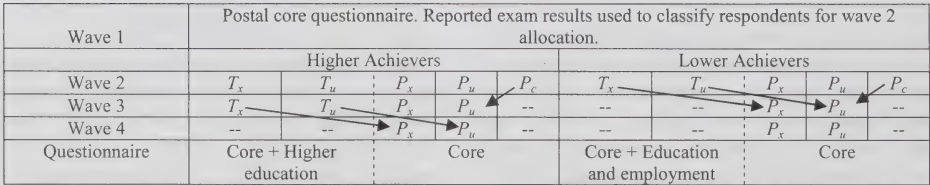
Wave 1 was a postal survey with telephone follow-up of nonrespondents after 4 mailings (initial questionnaire mailing and three reminders). Based on reported examination results, wave 1 respondents were classified as either ‘higher achievers’ (if they had obtained 5 General Certificate of Secondary Education examination passes at grades A\* to C) or ‘lower achievers’ otherwise. This led to around one-third of wave 1 respondents being classified as lower achievers. At wave 2 roughly one third of issued sample members were randomly selected for additional questions on particular topics and assigned to computer assisted telephone interviewing (CATI). In addition to the core questionnaire, telephone respondents were administered a module on decisions about entering higher education (for higher achievers) or on educational and employment

aspirations (for lower achievers). The remaining sample members were administered the core questionnaires by post. At wave 3 all lower achievers received the core mail questionnaire, although the telephone module continued to be carried for a third of higher achievers. At wave 4 all respondents were assigned to the core postal survey. Figure 1 illustrates the allocation to modes and incentives.

The core questionnaire remained mainly unchanged for the three experimental waves. Telephone respondents were asked the core questions before the additional modules. The core questionnaire was the same as the postal questionnaire, although some items were adapted for administration over the telephone. The average telephone interview took around 20 minutes. (The questionnaires and technical reports are available via the UK Data Archive in the appendices of the YCS User Guide at <http://www.data-archive.ac.uk/findingdata/snDescription.asp?sn=4571&key=YCS>.)

3.2 The incentives experiment

Facing growing concerns over declining response rates, an experiment was introduced in the second wave of cohort 10, to study the effect of incentive payments on response rates and nonresponse bias. A proportion of wave 1 respondents on both the postal and telephone surveys were sent a GBP5 voucher (approx. USD10 or EUR7), while the control groups received no such incentive. Additionally, in the postal survey the incentives were either unconditional (the incentive was sent with the initial mailing) or conditional (the voucher was promised in the original mailing, but only sent on receipt of a completed questionnaire). At waves 3 and 4, all incentives were paid unconditionally.



Notes:  $T$  = telephone,  $P$  = postal,  $x$  = control,  $u$  = unconditional incentive,  $c$  = conditional incentive. Arrows indicate changes in incentive treatment or mode allocation between waves.

Figure 1 Experimental design

3.3 Allocation of respondents to modes and incentive treatments

At wave 2, wave 1 respondents were randomly assigned to either telephone or postal mode. The allocation of incentive treatments was however done at the school level (randomised cluster assignment by mode). Each school represented in the sample was allocated to one telephone treatment (control or incentive) and independently allocated to one, potentially different, postal treatment (control, unconditional incentive or conditional incentive), so that all sample members from the same school approached in the same mode received the same incentive treatment.

The 4,712 wave 1 lower achiever respondents were stratified by identification number within school within Government Office Region and alternately allocated to telephone and postal treatments. We have excluded from the analysis 627 cases for which there was no valid telephone number on file, as those amongst this group who had been allocated to telephone mode were approached by post. Consequently, analysis of lower achievers is restricted to 2,097 approached by telephone and 1,988 approached by post.

A similar procedure was carried out for higher achievers, except that a larger proportion was allocated to postal treatment. There were 8,909 wave 1 higher achiever respondents of which 751 had no valid telephone number and are excluded from the analysis. After these exclusions there are 2,922 higher achievers allocated to telephone mode and 5,236 allocated to postal mode.

For the allocation of schools to incentive treatment groups, the schools containing telephone sample members (i.e. all schools apart from a few of the very smallest schools with fewer than five pupils in the sample) were stratified according to the ratio of lower to higher achievers in the

sample and randomly assigned to incentive treatments within strata. (The proportion of schools assigned to incentives was 1/2 if the ratio of lower to higher achievers in the sample was  $\geq 2$ ; 1/3 for  $1/2 \leq \text{ratio} < 2$  and 1/4 for all remaining schools.) The procedure was repeated for the allocation of schools in the postal treatment groups, where those selected for incentive treatment were randomly split into a conditional and an unconditional treatment group. (The proportions allocated to incentives were 2/3 if the ratio was  $\geq 2$ ; 1/3 for  $0 \leq \text{ratio} < 1/2$  and 1/6 for all other schools.) All estimates of significance presented in this text account for the clustered sampling design of the incentive experiment.

Table 1 shows the issued sample sizes at each wave for the different treatment and mode combinations, excluding cases of known ineligibility who had either moved abroad or died ( $n = 13$  at wave 2;  $n = 3$  at wave 3). Ineligible cases at wave 4 are not identified in the data, but the number is likely to be small. The analysis also excludes wave 1 respondents for whom no telephone number was known at the time of the allocation to modes for wave 2, as described above, and 117 higher achievers assigned to telephone mode at wave 2, who responded by post and were subsequently allocated to postal mode.

Table 1 also documents the observed wave-on-wave and cumulative response rates (AAPOR RR1). The rates are shown by achievement level and sequential mode/incentive combination. Wave-on-wave response rates for the higher achiever sample allocated to telephone control at wave 2 and moved to postal control at wave 4 (Col 1) were, for example, 76.82%, 69.13% and 72.21%. The issued numbers of cases declined from 2,075 to 1,101 across the three waves, because nonrespondents were not issued in subsequent waves.

Table 1 Conditional and cumulative response rates

Wave	Response Rate %	Higher Achievers					Lower Achievers				
		$T_x T_x P_x$	$T_u T_u P_u$	$P_x P_x P_x$	$P_u P_u P_u$	$P_c P_u P_u$	$T_x P_x P_x$	$T_u P_u P_u$	$P_x P_x P_x$	$P_u P_u P_u$	$P_c P_u P_u$
2	Conditional	76.82	80.91	78.23	86.45	82.32	65.21	70.41	64.93	75.00	71.35
	(Issued $n$ )	(2,075)	(728)	(3,262)	(1,004)	(967)	(1,282)	(811)	(807)	(608)	(569)
3	Conditional	69.13	73.17	73.07	81.91	81.36	59.09	70.93	63.36	71.93	70.20
	(Issued $n$ )	(1,594)	(589)	(2,551)	(868)	(794)	(836)	(571)	(524)	(456)	(406)
4	Cumulative	53.11	59.20	57.16	70.82	66.94	38.53	49.94	41.14	53.95	50.09
	Conditional	72.21	85.61	76.11	85.65	86.82	63.16	74.26	65.36	75.30	81.34
	(Issued $n$ )	(1,101)	(431)	(1,863)	(711)	(645)	(494)	(404)	(332)	(328)	(284)
	Cumulative	38.31	50.69	43.48	60.66	58.03	24.34	36.99	26.89	40.63	40.60

Notes: AAPOR Response Rate 1. Treatment groups are identified by  $T$  = telephone,  $P$  = postal,  $x$  = control,  $u$  = unconditional incentive,  $c$  = conditional incentive.  $T_x T_x P_x$  for example, refers to the sample allocated to telephone control at waves 2 and 3 and to postal control at wave 4. Conditional response rates are conditional on response at the previous wave. The base is the number of issued cases, which excludes previous wave nonrespondents and ineligible cases. Cumulative response rates are the percentage of wave 1 respondents remaining in the respondent sample. The base is the wave 2 number of issued cases, excluding three higher achievers ineligible at wave 3 ( $1 P_x P_x P_x$  and  $2 P_c P_u P_u$ ).

#### 4. Outcome measures and methods

The analysis is based on the sample of wave 1 respondents, since allocation to experimental treatments used information collected in the first wave and the corresponding characteristics of wave 1 nonrespondents are unknown. Our focus is therefore on attrition, conditional upon wave 1 response. This is the aspect of non-response that is particular to panel surveys, though of course it must be recognised that the characteristics of attrition are conditional on the characteristics of wave 1 response. The response rate at wave 1 (AAPOR RR1) was, 54.80%, excluding 5 cases of known ineligibility (Russell and Phelps 2001). This section describes the outcome measures and methods used to evaluate the hypotheses about the effects of incentives.

##### 4.1 Attrition

To test the effect of incentives on attrition, we estimated the probability of attrition as a function of the experimental design variables (*telephone mode*, *unconditional incentives*, *conditional incentives*, *lower achievers*) and their interactions. For each of the three experimental waves ( $t = 2, 3, 4$ ), we estimated a separate probit model of the probability of attrition, in each case using the wave 1 respondent sample as the base:

$$\begin{aligned} \Pr(\text{attrition}_{it}) = & F(\beta_{0t} + \beta_{1t}tel_i + \beta_{2t}unc_i + \beta_{3t}cond_i \\ & + \beta_{4t}la_i + \beta_{5t}tel_i * unc_i + \beta_{6t}la_i * tel_i \\ & + \beta_{7t}la_i * unc_i + \beta_{8t}la_i * cond_i \\ & + \beta_{9t}la_i * tel_i * unc_i + \varepsilon_i) \end{aligned} \quad (1)$$

where  $F$  is the probit link function. The estimated coefficients and standard errors from this model were then used to calculate predicted probabilities of attrition under different treatment conditions and to test for differences due to incentives.

##### 4.2 Item nonresponse

To test the effect of incentives on item nonresponse, we estimated count models of the number of items missing, using all non-filtered items from the core questionnaires in waves 2 ( $n = 44$ ), 3 ( $n = 48$ ) and 4 ( $n = 46$ ), where 'don't know' was counted as a missing value. We used the same specification of the predictors as for model (1) to estimate separate negative binomial regression models for each of the three experimental waves, conditional on response to the given wave. (Overdispersion meant that Poisson models did not fit the data: the  $P$ -value of the Likelihood Ratio test of equal mean and variance was 0.0000 for all three waves.)

The estimated coefficients and standard errors from these models were used to calculate predicted item nonresponse under different treatment conditions and to test for differences due to incentives.

##### 4.3 Attrition bias

To test the effect of incentives on attrition bias, we estimated the probability of attrition using model (1) but including wave 1 respondent characteristics and their interactions with the experimental design variables as predictors. We estimated separate probit models for attrition at each of the experimental waves ( $t = 2, 3, 4$ ) and for each characteristic, again using the wave 1 respondent sample as the base:

$$\begin{aligned} \Pr(\text{attrition}_{it}) = & F(\beta_{0t} + \beta_{1t}tel_i + \beta_{2t}unc_i + \beta_{3t}cond_i \\ & + \beta_{4t}la_i + \beta_{5t}tel_i * unc_i + \beta_{6t}la_i * tel_i \\ & + \beta_{7t}la_i * unc_i + \beta_{8t}la_i * cond_i \\ & + \beta_{9t}la_i * tel_i * unc_i + \beta_{10t}wlchar_i \\ & + \beta_{11t}wlchar_i * tel_i + \dots \\ & + \beta_{19t}wlchar_i * la_i * tel_i * unc_i + \varepsilon_i) \end{aligned} \quad (2)$$

where  $\beta_{11}$  to  $\beta_{19}$  are the coefficients for the interactions of the characteristic with the design variables. The coefficient for the respondent characteristic,  $\beta_{10}$ , provides information about the direction, magnitude and, in combination with its standard error, the significance of attrition bias for the postal, no incentive, higher achiever reference group. The interaction of the characteristic and the incentive indicators provide information about the change in attrition bias due to incentives. The significance of all interactions presented in this text was calculated following recommendations for nonlinear models by Norton, Wang and Ai (2004) using the command 'predictnl' in Stata version 9.

The characteristics tested were gender, school type, exam results, current activity (full-time education, employment, not in education, employment or training ("neet")), experience of unemployment, studying for vocational or academic qualifications, household composition (living with parent, partner, neither) and a set of attitudinal questions about employment and training. The wording of all questions is documented in Table 6. The characteristics chosen were those for which respondents and non-respondents could be expected to differ, based on previous studies of nonresponse in the YCS and other surveys and on nonresponse theories (Groves and Couper 1998; Lynn, Purdon, Hedges and McAleese 1994).



4.4 Reported results

Since coefficients from non-linear models cannot be interpreted substantively (Long 1997), we report predicted values based on the model estimates, rather than coefficients. Unless stated otherwise, the results are for the higher achiever group. To convey a sense of the magnitude of differences in outcomes across treatments, we report transformations of the predicted values, comparing each treatment with the comparison group, the higher achiever postal control.

5. Attrition, item nonresponse and attrition bias in the control groups

As a background to the evaluation of the effects of incentives, this section documents the extent of attrition, item nonresponse and attrition bias in the control groups, highlighting differences across waves, achievement levels and modes. Throughout the discussion the higher achiever postal no-incentive group is the reference category, with which all other treatments are compared.

5.1 Attrition

The predicted cumulative attrition rate among higher achievers allocated to the postal control group, increased from 21.77% in wave 2 to 56.53% in wave 4 (Table 2, Col 1). For *lower achievers* (Col 2), attrition rates in the postal control group were 61% higher at wave 2, but this difference decreased across waves to 29% at wave 4. The difference by achievement level was nonetheless significant in all three waves ( $P$ -value of  $\beta_4 = 0.0000$  for  $t = 2, 3, 4$ ). In *telephone mode* (Col 3), attrition rates in the control group were not significantly different at wave 2, but 9% higher at wave 3 ( $P$ -value of  $\beta_5 = 0.0034$  for  $t = 3$ ). This is contrary to findings from other studies, where nonresponse is generally lower in telephone mode due to the role of the interviewer in persuading respondents to take part in the survey. One possible reason for finding the opposite in this study is that for both the postal and CATI treatment groups, further attempts to obtain responses from initial non-respondents were made by telephone, so that only the postal group had a multi-mode treatment. Secondly, the burden of the wave 2 survey (measured by the interview length) was higher for the telephone respondents due to the additional modules, possibly leading to higher nonresponse at wave 3 than among the postal sample. The predicted cumulative response rates, which were the base for the calculation of percentage differences across treatment groups, are documented in the first three columns of Table 5.

5.2 Item nonresponse

The predicted number of missing items in the higher achiever postal control group was 2.89 at wave 2, falling to 1.75 at wave 4 (Table 3, Col 1). For *lower achievers* (Col. 2), the expected count for the control group was 21% higher at wave 2, with the gap increasing to 45% at wave 4. The differences by achievement level were significant in all three waves ( $P \leq 0.0001$  for  $\beta_4$ ,  $t = 2, 3, 4$ ). For *telephone mode* (Col 3), the predicted count was 4% lower at wave 2 and 12% lower at wave 3 ( $P = 0.0000$  for  $\beta_5$ ,  $t = 2, 3$ ), compared with postal mode. The predicted item non-response counts, used as the base for the calculations presented in Table 3, are documented in columns 4 to 6 of Table 5.

5.3 Attrition bias

Nonresponse in the higher achiever postal control group was differential for all of the domains tested (Table 4). The respondent samples significantly over-represented those living with their parents, in full-time education or studying for academic qualifications. Predicted attrition rates for those in full-time education in the higher achiever postal control group, for example, were 14% lower than for those not in full-time education at wave 2, with the difference increasing to 17% by wave 4 ( $P = 0.0000$  for  $\beta_{10}$ ,  $t = 2, 3, 4$ ). At the same time, the respondent samples under-represented males, those in secondary modern schools, with low or no exam results, who thought employers did not give young people the right training and that making plans for the future was a waste of time, those in full-time employment, those who had experienced unemployment and those who were studying for vocational qualifications. Bias was particularly strong with respect to qualifications. Those without any or with very low exam qualifications were around 50% more likely to have attrited from the sample by waves 3 and 4, compared to sample members with better qualifications. Similarly, those in full-time employment were 17% more likely than those not in employment (most of whom were still in education) to drop out at wave 2, with the difference increasing to 22% by wave 4.

Including background information used by the YCS for weighting (gender, school type, exam results and region) in the models did not affect the bias for any of the characteristics (in each wave and for each item, the  $P$ -value  $> 0.05$  from Wald tests of the equality of  $\beta_{10}$  estimated with and without background characteristics; not reported), except for bias with respect to qualifications, which was somewhat reduced when the background information was included.

The extent of attrition bias was mostly stable across waves, except for a few characteristics. In the higher achiever postal control sample, the under-representation of males significantly increased from waves 2 to 4 ( $P$ -value

from a Wald test of the equality of  $\beta_{10}$  across the two waves = 0.0295; not reported). For some of the other characteristics, the bias significantly decreased across waves. Nonresponse bias associated with attending a modern school fell between waves 3 and 4 and bias associated with not having any qualifications fell between waves 2 and 3 and again between waves 3 and 4.

For lower achievers there were few differences in the extent of attrition bias (not reported). Bias by gender, that is the difference in predicted nonresponse rates between males

and females, was 12% less than for higher achievers at wave 4 ( $P$ -value of the interaction between achievement level and gender was 0.0425 for  $t=4$ ), and bias by full-time employment was 4% less at wave 2 ( $P$ -value = 0.0269 for  $t=2$ ); bias according to attitudes on training provided by employers was 9% higher at wave 2 ( $P$ -value = 0.0056); bias according to whether studying for academic or vocational qualifications was higher at wave 2 (22% and 13%), 6% lower and 1% higher at wave 3, and lower at wave 4 (81% and 92%).

Table 2 Effect of incentives on attrition rates

Wave	Control groups			Incentives		Incentives by ability		Incentives by mode and ability	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$p_x^{ha}$	$\frac{p_x^{la} - p_x^{ha}}{p_x^{ha}}$	$\frac{T_x^{ha} - p_x^{ha}}{p_x^{ha}}$	$\frac{p_u^{ha} - p_x^{ha}}{p_x^{ha}}$	$\frac{(p_c^{ha} - p_x^{ha})/p_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(p_c^{la} - p_x^{la})/p_x^{la}}{(p_c^{ha} - p_x^{ha})/p_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	21.77	0.6112	0.0650	-0.3777	0.4966	0.7602	0.9763	0.4669	0.8471
( $P$ -Value)		(0.0000)	(0.2268)	(0.0000)	(0.0142)	(0.5085)	(0.4332)	(0.0556)	(0.6810)
3	42.86	0.3734	0.0941	-0.3191	0.7066	0.6820	0.6743	0.4074	1.4275
( $P$ -Value)		(0.0000)	(0.0034)	(0.0000)	(0.0592)	(0.7834)	(0.8287)	(0.0057)	(0.0861)
4	56.53	0.2933	-	-0.3040	0.8402	0.6179	0.7340	0.6597	0.8338
( $P$ -Value)		(0.0000)	-	(0.0000)	(0.2244)	(0.2535)	(0.8177)	(0.0911)	(0.9265)

Notes:  $P$  = postal,  $T$  = telephone,  $x$  = control,  $u$  = unconditional incentive,  $c$  = conditional incentive,  $ha$  = higher achievers,  $la$  = lower achievers. Column (1) shows the predicted attrition rate for the postal control higher achiever sample. The remaining columns show proportionate change in predicted rates.  $P$ -values of columns 2-4 represent standard errors of the main effects in the probit model; column 5 represents  $P$ -values from a Wald test of the equality of the coefficients for conditional and unconditional incentives; columns 7-9 represent  $P$ -values for the relevant interactions calculated using 'predictnl' in Stata version 9, according to Norton *et al.* (2004).

Table 3 Effect of incentives on item nonresponse (counts)

Wave	Control groups			Incentives		Incentives by ability		Incentives by mode and ability	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	$p_x^{ha}$	$\frac{p_x^{la} - p_x^{ha}}{p_x^{ha}}$	$\frac{T_x^{ha} - p_x^{ha}}{p_x^{ha}}$	$\frac{p_u^{ha} - p_x^{ha}}{p_x^{ha}}$	$\frac{(p_c^{ha} - p_x^{ha})/p_x^{ha}}{(4)^{ha}}$	$\frac{(4)^{la}}{(4)^{ha}}$	$\frac{(p_c^{la} - p_x^{la})/p_x^{la}}{(p_c^{ha} - p_x^{ha})/p_x^{ha}}$	$\frac{(T_u^{ha} - T_x^{ha})/T_x^{ha}}{(4)^{ha}}$	$\frac{(T_u^{la} - T_x^{la})/T_x^{la}}{(T_u^{ha} - T_x^{ha})/T_x^{ha}}$
2	2.89	0.2068	-0.9579	0.1008	1.3849	2.4825	0.6927	0.1820	-0.4094
( $P$ -Value)		(0.0005)	(0.0000)	(0.0173)	(0.4790)	(0.1308)	(0.6472)	(0.6251)	(0.9202)
3	2.54	0.3879	-0.8828	0.1660	1.5599	1.6788	1.2445	-0.9526	-0.1378
( $P$ -Value)		(0.0001)	(0.0000)	(0.0049)	(0.2372)	(0.4339)	(0.6796)	(0.0442)	(0.3890)
4	1.75	0.4533	-	0.0085	17.5491	16.8405	0.4621	13.8706	2.3073
( $P$ -Value)		(0.0013)	-	(0.9262)	(0.2133)	(0.4724)	(0.6481)	(0.5049)	(0.4530)

Notes:  $P$  = postal,  $T$  = telephone,  $x$  = control,  $u$  = unconditional incentive,  $c$  = conditional incentive,  $ha$  = higher achievers,  $la$  = lower achievers. Column (1) shows the predicted number of missing items of 44 non-branched items at wave 2, 48 at wave 3 and 46 at wave 4. The remaining columns show proportionate change in predicted item nonresponse counts.  $P$ -values of columns 2-4 represent standard errors of the exponentiated coefficients from the count model; column 5 represents  $P$ -values from a Wald test of the equality of the exponentiated coefficients for conditional and unconditional incentives; columns 7-9 represent  $P$ -values for the relevant interactions calculated using 'predictnl' in Stata version 9, according to Norton *et al.* (2004).

Table 4 Attrition bias (higher achiever postal control group)

	Wave 2	P-Value	Wave 3	P-Value	Wave 4	P-Value
Male	0.0807	(0.0000)	0.1330	(0.0000)	0.1474	(0.0000)
School type						
Comprehensive 16	0.0196	(0.2645)	0.0102	(0.6178)	0.0259	(0.2060)
Comprehensive 18	-0.0197	(0.1966)	-0.0138	(0.4444)	-0.0200	(0.2650)
Selective	-0.0188	(0.3661)	-0.0547	(0.0407)	-0.0213	(0.4577)
Modern	0.2310	(0.0001)	0.2423	(0.0004)	0.1597	(0.0261)
Independent	-0.0142	(0.4639)	0.0147	(0.5245)	-0.0068	(0.7756)
Exam results						
5+ grades A-C	-0.0977	(0.1778)	-0.0866	(0.3060)	-0.1795	(0.0320)
1-4 grades A-C	0.0831	(0.2857)	0.0721	(0.4298)	0.1696	(0.0606)
5+ grades D-G	0.0324	(0.8769)	0.0715	(0.7739)	0.1849	(0.4536)
1-4 grades D-G	-0.2177	(0.0000)	0.5714	(0.0000)	0.4347	(0.0000)
None	0.7826	(0.0000)	0.5716	(0.0000)	0.4348	(0.0000)
Attitudes						
Employers don't give training	0.0842	(0.0000)	0.0882	(0.0000)	0.0798	(0.0001)
Training more important than pay	0.0108	(0.4808)	-0.0070	(0.6979)	-0.0062	(0.7370)
Plans for future are a waste of time	0.0656	(0.0959)	0.1457	(0.0015)	0.1371	(0.0030)
Information about opportunities	0.0034	(0.8431)	-0.0204	(0.3266)	-0.0236	(0.2549)
Enough support planning future	0.0063	(0.6771)	0.0043	(0.8233)	-0.0105	(0.5848)
Current activity						
In full-time education	-0.1371	(0.0000)	-0.1462	(0.0000)	-0.1728	(0.0000)
In full-time employment	0.1661	(0.0003)	0.1983	(0.0001)	0.2201	(0.0000)
Neither in employment, education or training	0.0898	(0.1387)	0.1036	(0.1495)	0.1098	(0.1184)
ILO unemployed	0.0112	(0.6272)	0.0573	(0.0421)	0.0475	(0.0879)
Unemployed during past 12 months	0.0246	(0.4216)	0.0731	(0.0523)	0.0891	(0.0146)
Studying for academic qualifications	-0.1173	(0.0000)	-0.1351	(0.0000)	-0.1341	(0.0000)
Studying for vocational qualifications	0.0677	(0.0001)	0.0882	(0.0000)	0.0721	(0.0003)
Living arrangements						
Living with parent	-0.1348	(0.0111)	-0.1916	(0.0027)	-0.1033	(0.0986)
Living with partner	0.0904	(0.4457)	-0.0441	(0.7475)	-0.1042	(0.4525)

Notes: Predicted differences in attrition rates based on  $\hat{\beta}_{10,t}$ , i.e., prediction for each category compared to all residual categories. Each table entry is from a different model as explained in the text. *P*-values based upon estimated standard errors of the coefficient for the characteristic in the probit model.

Attrition bias in telephone mode was no different from postal mode, except for differential nonresponse by gender: the bias was 7% less at wave 2, 2% less at wave 3 and 1% more at wave 4 (*P*-value of the interaction between telephone mode and gender was  $\leq 0.002$  for  $t = 2, 3, 4$ ).

## 6. Evaluation of hypotheses

The evidence discussed here is summarised in Table 2 (effects of incentives on attrition), Table 3 (effects on item nonresponse), Table 4 (effects on attrition bias) and Table 5 (net effect on unit and item nonresponse).

*H1: Effects of incentives on attrition rate, attrition bias and item nonresponse.*

Incentives reduced attrition and increased item nonresponse but did not impact on attrition bias. Unconditional incentives reduced cumulative attrition in the postal higher achiever sample (Table 2, Col 4) by 38% (corresponding to an 8 percentage point difference) at wave 2, 32% at wave 3 and 30% at wave 4 (*P*-value of  $\beta_2 = 0.0000$  for  $t = 2, 3, 4$ ). At the same time, the incentive increased item nonresponse by 10% at wave 2 and 17% at wave 3 (*P*-value of  $\beta_2 \leq 0.05$

for  $t = 2, 3$ ), but had no effect at wave 4 (Table 3, Col 4). The difference across waves was however not significant (see *H2*).

Incentives had a proportionate effect on attrition across all respondent characteristics tested and therefore did not reduce attrition bias: the *P*-value of the interaction of unconditional incentives and respondent characteristics was  $> 0.05$  for all characteristics and waves (not reported). The exception was the proportion of pupils in 'modern' schools who were under-represented in all three waves. (Modern schools were the smallest category, representing only 2.8% of the wave 1 respondent sample.) Unconditional incentives reduced this bias by 60%, 47% and 78% at waves 2, 3 and 4 respectively (*P*-values of the interaction of incentives and modern school  $\leq 0.01$  for  $t = 2, 3, 4$ ).

Since incentives had a positive effect on unit response and a negative effect on item response, Table 5 documents the net effect on the amount of information collected in the survey. The benefits of incentives in terms of unit nonresponse clearly outweighed the cost in terms of item nonresponse. For each sample person issued at wave 2, the predicted unit and item response



rates for the postal higher achiever sample implied that by wave 4, 40% more valid items were collected with unconditional incentives compared to the control group. For lower achievers, 50% more information was collected with incentives. This is, however, a crude measure of the net effect of incentives, since in a multivariate analysis or for analyses of change, different patterns of missingness across items or across waves may lead to large numbers of cases being dropped by pairwise deletion.

*H2: Effects of incentives across waves.*

The effect on attrition decreased somewhat across waves, while the effects on item nonresponse and attrition bias were constant. Incentives reduced attrition by 38% at wave 2, 32% at wave 3 and 30% at wave 4 (Table 2, Col 4). The effects were similar at waves 2 and 3, but significantly different between waves 2 and 4 and between waves 3 and 4 ( $P$ -value from a Wald test of the equality of  $\beta_2$  across waves was  $\leq 0.05$ ). Although the relative effect of incentives decreased, the absolute effect increased across waves (-17 percentage points at wave 4, compared to -8 and -14 at waves 2 and 3, see Table 5). The effect of incentives on *item nonresponse* was not significantly different across waves ( $P$ -value of equality of  $\beta_2$  across waves was  $> 0.05$ ), although the predicted numbers of missing items fell across waves. Similarly, the effects of incentives on *attrition bias* did not differ across waves.

*H3: Conditional compared to unconditional incentives.*

Unconditional incentives had a greater effect in reducing attrition than conditional incentives, but similar effects on item nonresponse and attrition bias. For higher achievers, the conditional incentives used at wave 2 were only half as effective at reducing attrition as unconditional incentives (Table 2, Col 5) and the difference between the two conditions was significant ( $P$ -value from a Wald test of the equality of  $\beta_2$  and  $\beta_3$  was 0.0142). At the same time, conditional incentives increased *item nonresponse* by 38% more than unconditional incentives (Table 3, Col 5), but the difference was not significant. Conditional incentives somewhat reduced *attrition bias* for a single characteristic: sample members in the control group studying for vocational qualifications at wave 1 were 6.8% more likely to drop out than those not studying for vocational qualifications. With conditional incentives the difference was 6.4% ( $P$ -value of the interaction of conditional incentives with this characteristic was  $\leq 0.05$  for  $t = 2$ ).

*H4: Differential effects by mode.*

Incentives had more effect on attrition and item nonresponse in postal than telephone mode, but no

effect on attrition bias in either mode. In telephone mode, unconditional incentives had less than half the effect on *attrition* they had in postal mode for the higher achiever group (Table 2, Col 8). The difference was significant at wave 3 ( $P$ -value of the interaction between telephone mode and unconditional incentives was 0.0057) but not at wave 2. At wave 3, incentives increased *item nonresponse* 5% less in telephone mode than in postal mode ( $P$ -value of the interaction was 0.0442), but the difference at wave 2 was not significant. The lack of effect of unconditional incentives on *attrition* was no different across the two modes.

*H5: Effects of changes in mode or incentive treatment.*

Changing the incentive condition or mode did not have lasting effects. Changing the treatment from conditional to unconditional incentives had no lasting effect on either *attrition* or *item nonresponse* ( $P > 0.05$  from Wald tests of the equality of  $\beta_2$  and  $\beta_3$  for  $t = 3, 4$ ) and the effects after the change in treatment were similar to those for the sample allocated to unconditional incentives from the start (Tables 2 and 3, Col 5). Changing the survey mode from telephone to postal did not have a lasting effect on *attrition* or *item nonresponse* either ( $P$ -value of the interaction for telephone mode and unconditional incentives  $> 0.05$  at  $t = 4$ ) and the effects after the change in mode were no different from the effects for the sample allocated to postal unconditional incentives from the start (Tables 2 and 3, Col 8).

*H6: Differential effects by ability level.*

The effects of incentives were similar across achievement levels. Differences between achievement levels in the proportional effects of unconditional and conditional incentives on *attrition* and *item nonresponse*, were not significant (Cols 6 and 7 in Tables 2 and 3 report the  $P$ -values of the interactions of achievement level with each of the incentive treatments), since the absolute effects were comparable. Unconditional incentives, for example, reduced attrition at wave 2 by 8 percentage points among higher achievers and 10 percentage points among lower achievers. However, since the level of nonresponse in the control group was 61% higher for the lower achiever group, the similar absolute effect implied a smaller proportional effect of only 76% of the effect for higher achievers.

Similarly, the difference between modes was not differential by achievement (Tables 2 and 3, Col 9 report the  $P$ -values of the interaction between achievement level, unconditional incentives and telephone mode) and the lack of effect on attrition bias was no different for lower achievers (not reported).

Table 5 Net effect of incentives on unit and item response

		Predicted cumulative RR (%)			Predicted mean # INR			# valid items per unit issued at w2: incentive/control		
		w2	w3	w4	w2	w3	w4	w2	w3	w4
Higher	$P_x$	78.23	57.14	43.47	2.89	2.54	1.75	-	-	-
Achievers	$P_u$	86.45	70.82	60.66	3.19	2.96	1.77	1.097	1.228	1.395
	$P_c$	82.32	66.80	57.91	3.30	3.20	2.01	1.042	1.152	1.324
	$T_u$	76.82	53.11	38.31	0.12	0.30	1.61	-	-	-
	$T_u$	80.91	59.20	50.69	0.12	0.25	1.80	1.053	1.116	1.317
	$P_x$	64.93	41.14	26.89	3.49	3.52	2.54	-	-	-
Lower	$P_u$	75.00	53.95	40.63	4.37	4.51	2.91	1.130	1.282	1.498
	$P_c$	71.35	50.09	40.60	3.83	4.66	2.72	1.090	1.186	1.504
	$T_x$	65.21	38.53	24.34	0.50	3.48	2.35	-	-	-
	$T_u$	70.41	49.94	36.99	0.49	3.56	2.99	1.080	1.294	1.498

Notes: RR = response rate, INR = item nonresponse, # = number.  $T$  = telephone,  $P$  = postal,  $x$  = control,  $u$  = unconditional incentive,  $c$  = conditional incentive. Calculation based on 44 non-branched items at wave 2, 48 at wave 3 and 46 at wave 4. The number of valid items is calculated as  $RR_u(44- INR_2 + 48- INR_3 + 46- INR_4)$ .

Table 6 Question wording of items included in analysis of nonresponse bias

Variable	Question wording
Year 11 exam results	"Please tell us: a) Which GCSE subjects you studied in Years 10 and 11, b) Which GCSE subjects you have taken an exam in, c) Your GCSE results (do not record any re-sit results obtained in Year 11)."
Attitudes:	"Here are some things which people have said. We would like to know what you think. Please put a cross in one box for each statement: Agree, Disagree, Don't know."
ATT: employers	Agree: "Most employers don't give young people the right kind of training at work."
ATT: training/pay	Agree: "In looking for a job, I am more concerned to find one with training than one that pays the best."
ATT: plans	Agree: "I think that making plans for the future is a waste of time."
ATT: information	Agree: "I know how to find out about future work, training or education opportunities."
ATT: support	Agree: "I get enough support in planning my future."
Current activity:	"Please put a cross against one box to tell us your main activity at the moment: a) Out of work/unemployed, b) Modern Apprenticeship, National Traineeship, Youth Training or other government supported training, c) In a full-time job (over 30 hours a week), d) In a part-time job (if this is your <i>main</i> activity), e) In full-time education at school or college, f) Looking after home or family, g) Doing something else (please specify)."
In ft education	In full-time education.
In ft employment	In full-time employment.
NEET	Not in employment, education or training.
ILO unemployed	Unemployed and searching for job among economically active (YCS derived variable).
Unemployed	Unemployed in one or more months from April 1999 to March 2000: "We would also like to know what you have been doing over the past months. Please put a cross in one box for each month to show us what you were doing for all, or most of each month".
	Response options as for current activity, including 'On holiday'.
Studying (ac)	Yes: "At present, are you studying for GCSE, A/S or A-level qualifications?"
Studying (voc)	Yes: "At present, are you studying for any GNVQs (General National Vocational Qualifications)?" or "At present, are you studying for NVQ (National Vocational Qualification) or any other vocational or professional qualification including BTEC, City & Guilds or RSA qualifications?"
Household:	"Who lives in the same household as you? a) Father, b) Stepfather, c) Mother, d) Stepmother, e) Your own children, f) Brothers and sisters g) Other persons (please write in their relationship to you)."
Living with parent	Living with one or more of father, stepfather, mother or stepmother.
Living with partner	Living with boyfriend, girlfriend, husband, wife or partner.

## 7. Summary and discussion

This study has provided new evidence on the effects of continued incentive payments in a multi-mode panel study. We tested the effects of incentives on attrition, item nonresponse and attrition bias and whether these effects changed across waves. We also tested whether conditional

and unconditional incentives had similar effects, whether incentive effects were differential across modes and ability levels, and whether changes in the incentive treatment or mode had lasting impact on the effect of incentives in subsequent waves.

The findings showed that unconditional incentives significantly reduced attrition and, although they also

increased item nonresponse, the net effect on the amount of information collected by the survey was positive. Incentives had proportionate effects across a range of respondent characteristics and as a result did not impact on attrition bias in terms of those characteristics. Item nonresponse increased more with unconditional than conditional incentives, and more in postal than in telephone mode. Attrition bias was not affected by either incentive treatment in either mode. Across waves, incentives had a somewhat decreasing effect on attrition, but similar effects on item nonresponse. The lack of effect on attrition bias was also a constant across waves. Changes in incentive treatment from conditional to unconditional, and in mode from telephone to postal, did not affect outcomes at later waves.

The findings imply that respondent incentives are an effective means of maintaining sample sizes of a panel and ensuring its value in terms of efficiency of estimation and feasibility of subgroup analyses. Among lower achievers, fully 50% more information was collected during the three experimental waves, in terms of the number of valid items per case issued at the start. Incentives were safe, in the sense that increased response rates did not inadvertently increase nonresponse bias in terms of observed characteristics.

Changes in incentive treatment did not have lasting effect; however, in this study the only change implemented was an improvement for the respondent, from conditional to unconditional incentives. Expectations formed on the basis of previous incentive treatments may well mean that changes have lasting effect, if the change reduces the value of the incentive in the eyes of the respondent (see, Singer, Van Hoewyk and Maher 1998).

Incentives had no effect on attrition bias. We could however not evaluate the effect on bias of nonresponse at wave 1. Ideally, we would assess both the magnitude of bias due to nonresponse at wave 1 and due to subsequent attrition, and the effects of incentives on both. It is possible that nonresponse at wave 1 is more detrimental in terms of bias than later attrition, especially in studies such as the present one with low initial response rates. In this case, the effect of incentives on bias at wave 1 may be more important than any effect on bias caused by attrition. In addition, the discussion of the effects of incentives on attrition bias has focused entirely on observed characteristics and although incentives did not have differential effects in terms of these, they may nonetheless have differential effects in terms of unobserved factors. If this were the case, the use of respondent incentives could introduce sample selection bias in multivariate estimates, if the unobservables determining the responsiveness to incentives are correlated with outcomes measured by the survey (Kennedy 2003). For example, if responsiveness to incentives depends on time preferences for money and this

factor also determines the decision to leave further education and work instead, then models of the determinants of educational outcomes will lead to biased estimates.

Finally, there was little evidence that the respondent sample became less sensitive to incentives across waves as potentially less committed sample members dropped out. This finding is consistent with Laurie (2007), who reported that an increase in the value of an incentive in the British Household Panel Survey significantly increased response, even after 14 waves of the panel, with already high annual response rates of around 95% each year. Since previous studies have found that the effects of one-off incentives can carry over across waves (James 1997; Laurie and Lynn in press; Mack *et al.* 1998), a formal test of marginal effects of incentives would however require comparisons with a treatment group only offered an incentive at the first wave.

## Acknowledgements

We would like to thank Iain Noble at the Department for Children Schools and Families for facilitating the data and commenting on an earlier version of the paper, Tim Thair and Rory Fitzgerald for patient and helpful response to data queries, Noah Uhrig, Heather Laurie for comments and Mark Bryan for advice on the calculation of interactions in nonlinear models. The views expressed are those of the authors and not necessarily those of the Department for Education and Skills.

## References

- AAPOR (2006). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 4<sup>th</sup> Edition: American Association for Public Opinion Research.
- Armstrong, J.S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly*, 39, 111-116.
- Berk, M.L., Mathiowetz, N.A., Ward, E.P. and White, A.A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics*, 3, 449-457.
- Church, A.H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.
- Davern, M., Rockwood, T.H., Sherrod, R. and Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment. *Public Opinion Quarterly*, 67, 139-147.
- Downs, B. (1999). *Incentive Use in Panel Surveys*. Internal Census Bureau Memorandum. Washington, DC: US Census Bureau.
- Fox, R.J., Crask, M.R. and Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.



- Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161-178.
- Goyder, J. (1994). An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society*, 36, 360-366.
- Groves, R., Singer, E. and Corning, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.
- Groves, R.M., and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons, Inc.
- Hopkins, K.D., and Gullickson, A.R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education*, 61, 52-62.
- James, J.M., and Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly*, 54, September 1, 1990, 346-361.
- James, T.L. (1997). Results of wave 1 incentive experiment in the 1996 survey of income and program participation. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.
- Kennedy, P. (2003). *A Guide to Econometrics*. Oxford: Blackwell.
- Laurie, H. (2007). The effect of increasing financial incentives in a panel survey: An experiment on the british household panel survey, Wave 14. *ISER Working Paper No. 2007-05*. Colchester: University of Essex.
- Laurie, H., and Lynn, P. (in press). The use of respondent incentives on longitudinal surveys. In *Methodology of Longitudinal Surveys*, (Ed. P. Lynn), Chichester: Wiley.
- Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage.
- Lynn, P., Purdon, S., Hedges, B. and McAleese, I. (1994). England and Wales Youth Cohort Study: An Assessment of Alternative Weighting Strategies. Employment Department Research Series YCS No. 30. Sheffield: Employment Department.
- Mack, S., Huggins, V., Keathley, D. and Sundukchi, M. (1998). Do monetary incentives improve response rates in the survey of income and programme participation? *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.
- Martin, E., Abreu, D. and Winters, F. (2001). Money and motive: effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17, 267-284.
- Mason, R., Lesser, V. and Traugott, M.W. (2002). Effect of item nonresponse on nonresponse error and inference. In *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little), New York: John Wiley & Sons, Inc., 149-161.
- Norton, E.C., Wang, H. and Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4, 154-167.
- Rodgers, W. (2002). Size of incentive effects in a longitudinal study. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.
- Russell, N., and Phelps, A. (2001). Youth Cohort Study Cohort 10 Sweep 1: Technical Report. London: Taylor Nelson Sofres.
- Ryu, E., Couper, M.P. and Marans, R.W. (2006). Survey incentives: Cash vs. in-kind; face-to-face vs. mail; Response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18, 89-106.
- Shettle, C., and Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15, 231-250.
- Singer, E., Hoewyk, J.V., Gebler, N., Raghunathan, T. and McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15, 217-230.
- Singer, E., Van Hoewyk, J. and Maher, M.P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly*, 62, June 1, 1998, 152-164.
- Singer, E., Van Hoewyk, J. and Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, June 1, 2000, 171-188.
- Stratford, N., Simmonds, N. and Nicolaas, G. (2003). *National Travel Survey 2002: Report on Incentives Experiment*. London: National Centre for Social Research.
- Teisl, M.F., Roe, B. and Vayda, M. (2005). Incentive effects on response rates, data quality, and survey administration costs. *International Journal of Public Opinion Research*, 18, 364-373.
- Tzamourani, P., and Lynn, P. (1999). The Effect of Monetary Incentives on Data Quality - Results from the British Social Attitudes Survey 1998 Experiment. CREST Working Paper No. 73. Oxford: University of Oxford.
- Ward, R.K., Boggess, S., Selvavel, K. and McMahon, M.F. (2001). The use of targeted incentives to reluctant respondents on response rates and data quality. *Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria, VA: American Statistical Association.
- Willimack, D.K., Schuman, H., Pennell, B.-E. and Lepkowski, J.M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly*, 59, March 1, 1995, 78-92.
- Yu, J., and Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.



# Balancing sample design goals for the National Health and Nutrition Examination Survey

Leyla Mohadjer and Lester R. Curtin<sup>1</sup>

## Abstract

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States National Center for Health Statistics. A unique feature of NHANES is the administration of a complete medical examination for each respondent in the sample. To standardize administration, these examinations are carried out in mobile examination centers. The examination includes physical measurements, tests such as eye and dental examinations, and the collection of blood and urine specimens for laboratory testing. NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The major analytic goals of NHANES include estimating the number and percentage of persons in the U.S. population and in designated subgroups with selected diseases and risk factors. The sample design for NHANES must create a balance between the requirements for efficient annual and multiyear samples and the flexibility that allows changes in key design parameters to make the survey more responsive to the needs of the research and health policy communities. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the goals of NHANES.

**Key Words:** Multistage Sampling, Domain Sampling, Weighted Measure of Size, Mobile Examination Centers.

## 1. Introduction

The National Health and Nutrition Examination Survey (NHANES) is one of a series of health-related programs sponsored by the United States Centers for Disease Control and Prevention through its National Center for Health Statistics (NCHS). The NHANES surveys have been used to assess the health and nutritional status of the non-institutionalized civilian population of the United States for over 45 years. The data collected in NHANES are used to estimate the prevalence of major diseases and risk factors for diseases. The nutritional data from NHANES provide temporal monitoring for the nation with respect to such factors as diet, cholesterol, hypertension, iron deficiency, anemia, and obesity. NHANES has also been designed to assess the relationship between diet, health, and the environment so that nutritional assessments can be linked to such diseases as cardiovascular disease, diabetes, hypertension, and osteoporosis.

Data collection for NHANES includes at least three stages: a household screener, an interview, and a medical examination. The primary objective of the screener is to determine whether any household members are eligible for the interview and examination. The screener collects basic information on household composition and demographic characteristics. The interview collects household-, family-, and person-level data on demographic and socioeconomic background, health, and nutritional characteristics. Upon completion of the interview, respondents are asked to participate in a medical examination. To standardize administration and protocols, these examinations are carried out in a specially designed and equipped mobile examination center

(MEC). The examination includes physical measurements, tests such as eye and dental examinations, physiological measurements, and the collection of blood and urine specimens for laboratory testing. The NHANES website (<http://www.cdc.gov/nchs/nhanes.htm>) provides detailed information about the NHANES medical components.

The development of an efficient design has involved consideration of several design issues unique to NHANES in addition to the ones normally involved in survey samples. This paper is focused on the unique and challenging aspects of the NHANES design. However, it is helpful to provide an overall summary of the NHANES design, as given below, before discussing the unique features.

The NHANES sample represents the total non-institutionalized civilian population of the United States. Active military and institutionalized persons are not part of the population of inference. NHANES is not an equal probability design; sampling fractions are set to "over-sample" Mexican Americans (and Hispanics in the 2007 and beyond samples), black Americans, low-income white/other Americans, persons below age 20, and persons above age 60. A four-stage sample design is being used. The primary sampling units (PSUs), often referred to as stands, are selected from a frame of all U.S. counties. The PSUs are mostly single counties; in a few cases, adjacent counties are combined to keep PSUs above a certain minimum size. There are close to 3,000 PSUs in the NHANES sampling frame. NHANES PSUs are selected with probabilities proportionate to a measure of size (PPS). There are 15 stands in each annual sample.

The second sampling stage is area segments comprising Census blocks or combinations of blocks. Because PSUs

1. Leyla Mohadjer, Westat, 1650 Research Blvd., Rockville, Maryland, U.S.A. 20850. E-mail: [LeylaMohadjer@Westat.com](mailto:LeylaMohadjer@Westat.com); Lester R. Curtin, National Center for Health Statistics, 3311 Toledo Road, Hyattsville, Maryland, U.S.A. 20782. E-mail: [lrc2@cdc.gov](mailto:lrc2@cdc.gov).



vary in size, there is some variability in segment size and in the number of segments per PSU. Segments are formed with an average of about 150 households (or dwelling units) per segment. An average of about 5,000 segments are created within each PSU, and an average of 24 segments are sampled. The sample is designed to produce approximately equal sample sizes per PSU, and most PSUs have exactly 24 segments. The segments are also selected with PPS. The measures of size (MOS) of the segments, when combined with the subsampling rates used within the segments, provide approximately equal numbers of sample persons (SPs) per segment, although the relative variation in workload is greater among segments than among PSUs.

The third stage of sample selection consists of households and noninstitutional group quarters, such as dormitories. In a given PSU, following the selection of segments, all dwelling units (DUs) in the sampled segments are listed, and a subsample of households and group quarters within the DUs are designated for screening in order to identify potential SPs for interview and examination. SPs within the households or group quarters are the fourth stage of sample selection. All eligible members within a household are listed, and a subsample of individuals is selected. The subsampling rates for households within segments and for individuals within households are determined in advance. The combination of screening and differential sampling rates provides the increased sample size for those demographic subdomains of special interest (age, sex, race/ethnicity, and income). For example, in the 30 PSUs in which data were collected during the 2-year data cycle 2005-2006, 716 segments were selected and 26,529 households were selected for screening. After being screened for age, sex, and race/ethnicity composition and low-income status, 6,372 households had one or more individuals selected into the sample. A total of 12,862 individuals were selected, of whom 9,950 completed interviews and examinations.

The NHANES examination requires both highly specialized personnel and laboratory processing of collected specimens. As a result, examination components can be very costly to implement. To limit costs and reduce respondent burden, certain examination components are administered to only a subsample of MEC respondents. A single subsampling algorithm controls the amount of overlap among the various subsamples to allow analyses of correlations between various examinations and laboratory components. The SP's assignments to subsamples are fully determined before the SP arrives at the MEC.

The data collected in NHANES surveys have been extremely important in providing needed information about the health and nutritional status of the U.S. population. As a result, beginning with NHANES 1999, the survey has been

implemented as a continuous, ongoing, annual survey (Montaquila, Mohadjer and Khare 1998). It is critical to devote a lot of attention to the development and maintenance of an efficient sample design for such an important and complex survey. This paper discusses the challenges involved in designing and implementing a sample selection process that satisfies the multiple goals of NHANES. The paper focuses on the sample design used through 2006 (in response to emerging analytical requirements, some aspects of the sample design changed starting in 2007).

Section 2 outlines the major purposes and goals of the survey, followed by an overview of the major factors affecting the design given in Section 3. The unique features of the NHANES sample design are described in Section 4. Finally, Section 5 provides a brief summary of the paper.

## 2. Major purposes and goals of NHANES

NHANES is an ongoing annual health survey of the noninstitutionalized civilian population of the United States. The main objectives of NHANES are to (1) estimate the national prevalence of selected diseases and risk factors; (2) estimate national population reference distributions of selected health parameters and environmental contaminants; (3) document and investigate reasons for secular trends in selected diseases and risk factors; (4) contribute to the understanding of disease etiology; (5) investigate the natural history of selected diseases; (6) study the relationship between diet, nutrition, environment, genetics, and health; and (7) explore emerging public health issues.

## 3. Major factors affecting sample design

As mentioned above, a unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the design needs to produce efficient sample sizes for a large number of subdomains of the general population. Many health and nutritional characteristics differ considerably by age, sex, and race/ethnicity and are also affected by income status. As a result, most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Therefore, the survey is designed to produce efficient sample sizes for a very large number of subdomains of the U.S. population.

In general, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. More specifically, the NHANES design attempts to (1) obtain prespecified self-weighting sample sizes for a set of about 75 predesignated subdomains; (2)

produce sample sizes per PSU that will result in an efficient workload for the interview and examination staff at the MEC; (3) design samples that are likely to achieve high response rates; (4) be as cost effective as possible; (5) produce efficient annual samples; (6) allow for accumulation of samples, especially for rare subdomains or rare diseases over time; and (7) be flexible to allow changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

In the remainder of this section, we provide brief summaries of how each of these seven goals affects the design and implementation of NHANES.

*NHANES subdomains* - The sample design for NHANES meets a prespecified level of precision for cross-sectional data and comparisons over time for a set of predesignated subdomains. Specifically, 77 sampling domains (in the 2006 sample) are defined by race/ethnicity, sex, age, income, and pregnancy status. The sample includes oversamples of blacks, Mexicans, the very young, adolescents, the elderly, pregnant women, and the low-income population.

When estimates of universe totals for the entire population are considered to be of the greatest importance, then the best available estimate of the total population is used as an MOS in the sample selection process. For NHANES, where the interest is in subdomains of the total population, an alternative MOS is needed to improve the accuracy of the estimates and provide better control of the sample size. Section 4 describes the MOS used for sampling PSUs and segments in NHANES.

The objective of oversampling (using differential probabilities of selection) is to achieve a sample containing proportionately more members of certain population subdomains than there are in the population. The goal is to obtain adequate sample sizes to make inferences for subdomains representing relatively small proportions of the total universe of interest and to do it in such a way as to minimize variances for the budget available for the survey. Different oversampling strategies are used depending on the domains of interest. For example, oversampling of the minority subpopulations is accomplished through stratifying geographic areas by concentration of these minority groups and selecting segments in high-density areas at a higher rate. On the other hand, a large screening sample may be required to oversample persons within specific age categories. The subsection on Cost Ratios below describes why oversampling procedures used in NHANES are different from those commonly used in many area frame sample surveys.

*Workload for mobile examination centers (MECs)* - The MEC consists of four specially designed and equipped trailers and contains all of the medical equipment. Each trailer is approximately 45 feet long and 10 feet wide.

Detachable truck tractors drive the trailers from one location to another. MECs travel to survey locations throughout the country. The trailers are set up side by side and connected by enclosed passageways. The area in the MEC is divided into rooms to allow privacy during the examinations and interviews. The examination includes a variety of physical and dental assessments and measurements, laboratory tests, and health interviews.

Because of the logistical issues related to the traveling MECs, the sample size in each sampled location must be derived ahead of time and considered fixed so that field operations can be scheduled in an efficient and manageable way. Also, it is necessary to establish a firm time schedule for each stand so that appointments can be made for examinations. It is not possible to change the time schedule since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

*Response rates* - Achieving high response rates is a concern for practically every sample survey. With NHANES, this is a particular challenge because of the extensive nature of the interviews and examinations. Remunerations have been used in NHANES as a means of improving response rates. In addition, NHANES has an extensive outreach program that includes contacts with local organizations and individuals to gain cooperation, as well as local media coverage to reach as many SPs as possible. As a sample design issue, one approach that has been proven to favorably affect response rates is selecting larger sample sizes within sampled households. One of the factors thought to be responsible for the increased response rates in multiple-SP households is that each person is given remuneration for his or her time and participation, and it is generally more convenient for household members to come to the MEC at the same time. Table 1 shows the examination response rates for SPs coming from households where only one person was selected compared to the response rates for SPs coming from multiple-SP households. As the table indicates, response rates increase by about 4 to 7 percent depending on the type of household.

NHANES is, therefore, designed to maximize the number of SPs per household. Such an approach is feasible for studies like NHANES, where the sample is composed of a large number of subdomains. That is, the effect of within-household clustering is not a large concern for NHANES because most analyses are done within age-sex-specific subdomains (or some limited groups of subdomains) and there is generally little within-household clustering at the subdomain level. The average number of SPs selected per household (in households where at least one SP was selected) within the defined sampling domains ranges from 1 to 1.24 in the 1999-2006 sample. Combining the domains down to 12 to 15 domains by collapsing over age and/or



race/ethnicity will result in average numbers ranging from 1.01 to 1.37 SPs per household. Therefore, some level of clustering is present to the extent that collapsed domains are used for analysis. Note that the SP sample is basically used for SP-level analysis (e.g., health and nutrition statistics). The clustering of SPs is, of course, higher at the family and household levels. However, household- or family-level variables are used for such analysis (e.g., household dust levels, family income, or insurance). Refer to Curtin and Mohadjer (2008) for a discussion of the impact of clustering, and unequal probabilities of selection of subdomains, on the precision levels of various estimates.

**Table 1**  
Examination response rates by number of SPs in household, by household type, in 1999-2006 NHANES sample

Household type	Number of SPs selected per household		Response rate (%)	
	One SP	Two or more SPs	One SP	Two or more SPs
Black/Mexican	4,892	20,222	76.5	82.3
Other low-income <sup>1</sup>	1,362	3,349	77.6	84.5
Other non-low-income	5,597	15,508	68.8	72.6

<sup>1</sup> The Other group includes all SPs who are not Black or Mexican.  
The low-income threshold is set at 130 percent of poverty.

**Cost ratios** - The field data collection cost in area survey samples includes the cost of listing DUs, screening households to locate eligible respondents, and conducting the interview to collect data. In NHANES, the data collection phase includes both the household interview and the MEC examination. NHANES requires highly specialized medical equipment, personnel, and laboratory processing. As a result, the cost of an examination is very high compared to other costs in the survey. In fact, the cost of listing and screening is only about 3 to 4 percent of the cost of interviewing and examination. This cost ratio (the cost of interviewing and examination relative to the cost of listing and screening) greatly affects the design of NHANES.

As mentioned above, many of the predesignated subdomains of NHANES require some method of oversampling to achieve the required sample sizes. For the minority populations, substantial reductions in screening are possible with oversampling of highly concentrated minority areas. In general, an optimum design is developed by ascertaining the effect on cost and variance of alternative sampling procedures and choosing the one that minimizes the variance for a fixed cost. In the evaluation of trade-offs between cost and variance, suppose that a particular oversampling strategy reduces the number of households to be listed and screened while increasing the variance for most statistics. The savings brought about by the reduction

in cost of listing and screening could be used to increase the size of the sample and thereby lower the variance. However, in NHANES, listing and screening a household is only a very small fraction of the cost, and thus, it takes very large savings in listing and screening costs to justify a moderate increase in variance. As a result, the oversampling procedures established for the survey reflect the NHANES cost ratio and are different from those of typical area surveys.

**Annual and multiyear samples** - To facilitate potential linkage with other large-scale surveys, to retain flexibility in the sample design, and to allow for the production of annual estimates for broad subdomains, NHANES became a continuous, annual survey starting in 1999. The travel requirements for nationally representative annual samples in the United States are challenging. Three MECs – two of which are stationed at PSUs and one of which is traveling at any given time – work on a very carefully designed schedule to meet the design requirements of the study.

The ability to make meaningful inferences from any survey is affected by both the precision of the estimates themselves and the precision of the variances of the estimates used in the analysis. One of the main limitations of an NHANES annual sample is the small number of PSUs (15 per year), which results in a small number of degrees of freedom for both estimation and analysis and thus design-based variance estimates that are relatively imprecise. Additionally, the effective sample sizes for most subdomains are too small in annual samples. Most subdomain analyses will need to accumulate a number of annual samples to provide both precision and statistical power for comparisons. The procedures for combining years of the survey must be relatively simple, and appropriate for commercial software packages, to maximize the usefulness to the wide variety of users of the NHANES data. Thus, it is critical to employ a sample design that allows efficient accumulation of the annual samples across years.

**Flexible design** - A critical objective of NHANES is to explore emerging public health issues. The survey needs to be flexible and able to adapt to changing requirements and new challenges. Thus, the sample design must balance the need for efficient subdomain samples with the flexibility needed to make changes in key parameters. To date, the current NHANES design has been able to incorporate some changes in subdomain definitions and sampling rates when these changes have been made after the selection of PSUs. However, in extreme circumstances, substantial changes in subdomain definitions or sample size requirements would necessitate the selection of a new PSU sample.



#### 4. Unique features of the NHANES design

The factors described in Section 3 have played major roles in the development of the sample design and have resulted in some design features that are unique to NHANES. The unique features of the sample design include (1) weighted PSU and segment MOS; (2) efficient annual and multiyear samples; (3) maximized number of SPs per household; (4) controlled sample sizes for PSUs; (5) sequential release of the PSU sample; (6) special methods to deal with deterioration of the efficiency of the optimum design over time; and (7) special methods to reduce the risk of data disclosure through geographic identification.

The following paragraphs briefly describe these unique features of the NHANES design.

*Clustering and measures of size (MOS)* - In NHANES, the sample size must be large enough to produce an efficient workload for each PSU, considering the time and the cost involved in moving a MEC between survey locations and the time required to set up and break down the MECs for travel. Experience gained in earlier NHANES surveys has indicated that an average of 340 examined SPs is an approximately optimum number that provides the maximum number of PSUs while keeping the sample size in each area large enough to justify the costs associated with moving the MECs. In addition, the PSUs for NHANES are typically defined as individual counties to reduce the amount of travel necessary for respondents to visit a MEC, and thereby increase the likelihood of achieving high response rates.

The NHANES sample is designed to yield a self-weighting sample for each sampling subdomain while producing an efficient workload in each PSU. PSUs and segments are selected with probabilities proportionate to a weighted MOS, reflecting the PSU population in subdomains of interest. The selection probability of a PSU determines the maximum rate at which persons residing in that particular PSU can be selected. Refer to *Vital and Health Statistics, Series 2, No. 113, September 1992, CDC/NCHS*, available at <http://www.cdc.gov/nchs/products/pubs/pubd/series/sr02/120-101/120-101.htm>, for a description of the MOS used in NHANES.

*Annual and multiyear samples and stratification* - One way to achieve nationally representative annual samples is to select an independent sample of PSUs each year. Because of the limited number of NHANES PSUs and the fact that PSUs are selected proportionate to size, this approach would be likely to lead to substantial overlap in PSUs from year to year. Sample overlap, even at the PSU level, could lead to loss of precision in survey estimates when survey years are combined (due to increased clustering of the sample). Thus, rather than sampling PSUs independently each year, the approach in NHANES has been to select a 6-year sample,

from a nested structure of major and minor strata (as described below), and then allocate one PSU from each major stratum to each year. This nested structure for the 6-year sample avoids overlap of non-self-representing PSUs during the 6 years.

The design for the NHANES 6-year sample is a stratified two-PSU-per-stratum design and has been developed with the primary goal of efficiency for the 6-year sample, as well as efficient multiyear samples. The stratification scheme is designed to ensure that the PSUs comprising the annual and multiyear samples are distributed evenly in terms of geography and certain population characteristics.

The NHANES design (through 2006) included 18 self-representing PSUs. These PSUs ranged from those that were self-representing for the annual samples to those that were self-representing for 3-year or 6-year samples. These PSUs were assigned such that each year had an equal number of self-representing PSUs, with 3-year self-representing PSUs being 3 years apart. The non-self-representing PSUs were stratified into 12 major strata, defined based on geography and the metropolitan statistical area status of the PSUs. Seventy-two minor strata were defined based on the demographics of the PSUs. The minor strata were constructed to be of equal size to the extent possible (in terms of total MOS). The variables used to form the boundaries of the minor strata were minority status and the percentages of the population below poverty level. Each major stratum included six minor strata, and one PSU was selected from each of these final strata. Within each major stratum, minor strata were paired to create pseudo-strata. Each pair was randomly assigned to the study 3 years apart. The assignment of the pairs to the particular sets of study years and the assignment of the study years within the pair were random within the first major stratum, and all other major strata followed the same pattern.

This stratification scheme resulted in a sample of 72 non-self-representing PSUs that produces efficient annual and multiyear estimates without compromising the efficiency of the 6-year estimates. The 6-year sample has a one-PSU-per-minor-stratum design (or a two-PSU-per-pseudo-stratum design), and each annual sample has a one-PSU-per-major-stratum design. In addition, this design allows for the flexibility needed to address changes in the sample requirements (if a new sample needs to be selected), since the first 3 years of the sample follow a one-PSU-per pseudo-stratum design.

*Maximized number of SPs per household* - After the sample of screened households is identified, a sample of persons to be interviewed and examined from individual households is selected. All eligible members within a household are listed, and a subsample of individuals is selected based on sex, age, race/ethnicity, and income (all

pregnant women are selected with certainty). SPs are selected at rates established to ensure that the target sample sizes by subdomains will be achieved.

The sample of SPs is selected in a way that maximizes the average number of SPs per household in order to increase the overall response rate in the survey. If independent random selections are made for the subdomains, in most cases only one person in a household would be selected and the average sample size per household would be quite low, not much above 1. Therefore, instead of unrestricted randomization, a pseudo-random procedure is used that maximizes the number of SPs per households. Refer to Waksberg and Mohadjer (1991) for a description of the approach.

*Controlled sample sizes per PSU* - The sample size in each PSU (stand) that is actually generated from a self-weighting sample in each domain is based on a number of assumptions such as the age and race/ethnicity distribution in the PSU. These assumptions hold only approximately. Once the sample sizes have been calculated, they are treated as quotas, and the number of SPs in each stand is forced to adhere closely to the quota. The reason for this procedure is to have a manageable and efficient field operation. It is necessary to establish a firm, and fixed, time schedule for each stand so that appointments can be made for SP examinations. The time schedule obviously takes into account the expected number of SPs in each stand. As mentioned above, it is difficult to change the time schedule for a stand since it must be coordinated with the MEC's visits to other stands, which are also planned in advance.

There is no way of knowing in advance whether the assigned quota for a particular stand is lower or higher than what would arise from self-weighting samples within the various domains. Part of the reason for the uncertainty is that the MOS used for sample selection is based on the latest decennial Census and may not be quite up to date. The issue is further complicated by variations in response rates from stand to stand, as well as sampling variation in the number of identified SPs. Consequently, it is necessary to use a sample selection procedure that can produce samples that are either somewhat larger or somewhat smaller than those arising from the application of the self-weighting sampling rates.

*Sequential release of the sample in each stand* - To accomplish the above objective, an initial sample is selected in each stand that uses sampling rates 50 percent larger than those required to attain the target sample sizes in each domain. Each stand's initial sample is then divided into a group of subsamples. Each subsample is a systematic subsample of the initial sample, with the households sequenced by segment number and a temporary, geographically based sequence number prior to subsampling. Thus, each

subsample cuts across all segments, except when limited by sample size.

As a general rule, the 50 percent subsample (*i.e.*, subsample A) is released to the interviewers first. The yield from this subsample is monitored and used to project estimates of the total number of SPs expected when screening of this subsample has been completed. Based on these figures, additional subsamples are released as needed. The sample is monitored on a daily basis to determine whether additional subsample releases are required.

The one operational problem with the procedure for monitoring the sample yield is that it cannot completely control the subdomain sample sizes. The distribution of subdomains differs, to some extent, from the expected numbers based on the most recent Census data (used to derive the sampling rates). Experience with NHANES indicates that some population changes that will affect the sample sizes can be expected. Other factors that affect subdomain sample yield are patterns of nonresponse and undercoverage in stands. One option to correct the shortfall (or overage) in subdomain sample sizes is to change the sampling rates in future stands. However, such changes will increase heterogeneity in sample weights, thus adversely affecting the precision of the subdomain estimates, and are not advisable except under extreme circumstances.

*Dealing with deterioration of the efficiency of the optimum design over time in a tightly controlled sample* - The usual practice in area samples is to list all households in the sample segments and apply a prespecified sampling rate to the listed households. This approach gives all households the desired probabilities of selection. For example, if the sampling rate is 50 percent, then one-half of the housing units listed in the segments will be included in the sample. If the number of housing units has tripled due to new construction (*i.e.*, housing units built since the most recent decennial Census), the same sampling rate will produce three times as many interviews and examinations as the number originally expected. Such dramatic changes in the segment size are expected when the data collection period is several years after the most recent decennial Census for which data files are available.

For NHANES, highly variable sample sizes are not feasible because of the scheduling requirements of the MECs. Subsampling within PSUs, in an effort to obtain equal sample sizes across PSUs, is not recommended either, because it will introduce unequal weighting factors that would reduce the efficiency of the sample.

NHANES has used two procedures to update the segment MOS: (1) creation of new construction segments and (2) two-phase sampling to update the MOS. A third approach under consideration involves using purchased commercial address listings to update the MOS in a two-phase sample design.



Under the new construction approach (Bell, Mohadjer, Montaquila and Rizzo 1999), newly constructed units are excluded from area segments and new segments are created based on U.S. Census Bureau information on permits issued for new construction since the most recent decennial Census. New construction segments comprise clusters of building permits issued during one or several adjoining months by a building permit office. Census Bureau files from the Building Permits Survey are used as sources of the data on the number of residential building permits issued by the building permit offices.

Two-phase sampling is used in a number of statistical applications. One of the applications of two-phase sampling is to update a sampling frame when the sample is to be selected with respect to an MOS but a reliable estimate of the MOS is not available. With this approach, a larger sample of units (segments, in the case of NHANES) is selected. An updated value of MOS is then collected for this larger sample (also referred to as the first-phase sample). The final sample of units (segments) is selected from the first-phase sample using the updated MOS.

Starting in 2000, the NHANES segment MOS has been updated (for stands for which such updating seemed necessary) using a two-phase sampling procedure (Montaquila, Bell, Mohadjer and Rizzo 1999). In these cases, listers travel to the stand to obtain a count of the number of DUs in each segment in the first-phase sample. Using the listers' counts, an updated MOS that reflects the ratio of the actual number of DUs to the expected number of DUs is calculated for each first-phase segment. The final sample of segments is then selected by subsampling from the first-phase segments using the updated MOS.

*Risk of data disclosure through geographic identification* - In today's world, confidentiality concerns and the risk of data disclosure present real challenges to survey sponsors. The ability to identify survey respondents, either through unique combinations available on a single data file or by linking different databases, is of great concern. This is particularly true for NHANES, because of the extensive amount of sensitive data collected on each SP and the small number of PSUs in the sample. Therefore, NHANES evaluates the risk of disclosure on two fronts: geographic disclosure and disclosure from individual characteristics. Various methods (limited or suppressed data release) are used by NCHS to mask the individual characteristics that have a high risk of identifying individuals in the NHANES sample. Sensitive, limited, or non-released data items are available through a Research Data Center. At this time, only national estimates can be produced from publicly available data files; detailed geographic analyses must be done in the Research Data Center.

Although only national estimates can be produced, the direct estimation of sampling errors for those national estimates requires the release of design variables such as stratum and PSU identifiers. Typically, these variables indicate that a group of SPs are all in the same county but do not identify that county. Geographic disclosure is of a particular concern because (1) NHANES has a small number of PSUs, (2) PSUs are limited in geography to one county, and (3) an extensive amount of outreach activity is conducted within each PSU to improve response rates. The outreach program includes contacting various organizations and individuals at each stand to seek their support and using media (newspapers, television, and radio) to reach as many SPs as possible. It is therefore relatively easy to determine the counties in the NHANES sample. The racial/ethnic composition of a county, along with metropolitan/non-metropolitan status, is enough information to correctly match a list of known counties with groups identified as a county cluster on the public data file. To limit geographic disclosure, probabilistic record swapping methods are used at the second stage of sampling (segment swapping) to create masked variance units. The goal is to reduce the risk of identifying individuals by masking their location. Refer to Park, Dohrmann, Montaquila, Mohadjer and Curtin (2006) for a description of the swapping procedures applied to the NHANES sample.

## 5. Summary and conclusions

A unique feature of NHANES is the complete medical examination carried out in the MECs. In addition, the survey is designed to produce efficient sample sizes for a large number of subdomains of the U.S. population, since most analyses of NHANES data are conducted for defined age categories within various socioeconomic subgroups of the population. Thus, the sample design for NHANES must create a balance between the requirements for efficient subdomain samples and the need for an efficient workload for examination staff at the MEC, while keeping response rates as high as possible. In addition, the design must be as cost effective as possible, produce efficient annual samples, and allow for accumulation of samples for rare subdomains or rare diseases over time. Furthermore, the design must be flexible to allow for changes in key parameters, including sampling domains, and sampling rates to respond to emerging health issues.

The above requirements result in a very complex design with some design features that are unique to NHANES. In particular, the current sample is designed to produce efficient annual and multiyear samples. NHANES uses weighted PSU and segment MOS to yield self-weighting



samples for each subdomain, while producing an efficient workload in each PSU. Once the sample sizes are calculated, they are treated as quotas. The sample sizes are strictly controlled in each PSU to create a manageable and efficient field operation. A very large screening sample is used to oversample most of the age and income subdomains, and oversampling of highly concentrated areas is used for some of the very rare minority subdomains. The sample of SPs is selected using a pseudo-random procedure to maximize the average number of SPs per household because it has appeared to increase the overall response rate in previous surveys.

The challenges described in this paper are focused on the main aspects of the NHANES. There remain many other features unique to NHANES that analysts must take into account when analyzing data from the survey. For example, not only are there very few PSUs in each annual sample, but data collected within these PSUs are not randomly collected across the seasons. In particular, if there is a seasonal by geographic region interaction for a variable of interest, the current NHANES design will not be able to estimate it. Because of the small number of PSUs in each data release cycle, any contextual data linkage at the geographic level must be done in the NCHS Research Data Center. Because of the many subsamples within NHANES, special care must be taken to use the appropriate subsample weight; for example, estimates for undiagnosed diabetes must use the special fasting weight.

To facilitate the efficient use of MECs for data collection, there has been no attempt to randomly allocate the sample of PSUs across time in annual samples. However, the time dimension plays a major role in some health indicators, such as nutrition. Furthermore, analysis of nutrition data may also be affected by the complex nature of the design and data collection. Special sample weights constructed for the 2 days of the 24-hour recall data account for variation in the number of examinations by day of the week. A web-based

tutorial is now being developed to provide assistance in the analysis of NHANES nutrition data. A general tutorial for design-based analysis of NHANES data can be found at <http://www.cdc.gov/nchs/tutorials/>.

## Acknowledgements

The authors are grateful to the associate editor and the referees for their helpful comments and suggestions, which have greatly improved the paper.

## References

- Bell, B., Mohadjer, L., Montaquila, J. and Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Curtin, L.R., and Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, to appear.
- Montaquila, J., Bell, B., Mohadjer, L. and Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Montaquila, J., Mohadjer, L. and Khare, M. (1998). The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Park, I., Dohrmann, S., Montaquila, J., Mohadjer, L. and Curtin, L.R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Waksberg, J., and Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

# JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents

### Volume 23, No. 4, 2007

Non-Bayesian Multiple Imputation Jan F. Bjørnstad .....	433
Discussion	
Ray Chambers .....	453
Ralf T. Münnich .....	455
Chris Skinner .....	463
Seppo Laaksonen .....	467
Daniel Thorburn .....	477
Rejoinder Jan F. Bjørnstad .....	485
The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial Attitude Surveys Volker Stocké .....	493
The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease with Age Bärbel Knäuper, Norbert Schwarz, Denise Park and Andreas Fritsch .....	515
Dependent Interviewing and Seam Effects in Work History Data Annette Jäckle and Peter Lynn .....	529
The Influence of End-Users on the Temporal Consistency of an International Statistical Process: The Case of Tropical Forest Statistics Alan Grainger .....	553
Book and Software Reviews .....	593
Editorial Collaborators .....	603

## Contents

### Volume 24, No. 1, 2008

The Morris Hansen Lecture 2004 Bridging the Gap: Moving to the 1997 Standards for Collecting Data on Race and Ethnicity Jennifer H. Madans .....	1
Discussion	
Clyde Tucker .....	13
Robert B. Hill .....	17
A Model-free Bayes Analysis of Stratification, Clustering and Regression in Finite Population Survey Data Murray Aitkin .....	21
On the Covariance Between Related Horvitz-Thompson Estimators John Wood .....	53
Probability Based Estimation Theory for Respondent Driven Sampling Erik Volz and Douglas D. Heckathorn .....	79
Survey Experiences and Later Survey Attitudes, Intents and Behaviour Lars R. Bergman and Robert Brage .....	99
Temporal Disaggregation and Seasonal Adjustment Tommaso Proietti and Filippo Moauro .....	115
New Approaches to Creating Data for Economic Geographers Matthew Freedman, Julia Lane and Marc Roemer .....	133
Book and Software Reviews .....	157

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

## Volume 35, No. 4, December/décembre 2007

Bent JØRGENSEN & Peter X.K. SONG Stationary state space models for longitudinal data.....	461
Petra BŮŽKOVÁ & Thomas LUMLEY Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables .....	485
Jae Kwang KIM & Jay J. KIM Nonresponse weighting adjustment using estimated response probability .....	501
Jinhong YOU & Gemai CHEN On inference for a semiparametric partially linear regression model with serially correlated errors.....	515
Alexander R. DE LEON & K.C. CARRIÈRE General mixed-data model: extension of general location and grouped continuous models .....	533
Cinzia CAROTA A family of power-divergence diagnostics for goodness-of-fit .....	549
Chiung-Yu HUANG, Jing QIN & Fei ZOU Empirical likelihood-based inference for genetic mixture models .....	563
Xingwei TONG, Chao ZHU & Jianguo SUN Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments.....	575
Jesse FREY Distribution-free statistical intervals via ranked-set sampling.....	585
Heng LIAN Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces.....	597
Forthcoming papers/Articles à paraître.....	613
Volume 36 (2008): Subscription rates/Frais d'abonnement .....	614



# GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 32, No. 2 and onward) as a guide and note particularly the points below. Articles must be submitted in machine-readable form, preferably in Word. A paper copy may be required for formulas and figures.

## 1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ( $8\frac{1}{2} \times 11$  inch), one side only, entirely double spaced with margins of at least  $1\frac{1}{2}$  inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

## 2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

## 3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, *etc.*
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (*e.g.*, w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

## 4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

## 5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, *e.g.*, Cochran (1977, page 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

## 6. Short Notes

- 6.1 Documents submitted for the short notes section must have a maximum of 3,000 words.

# DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, priez d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 32, N° 2) et de noter les points ci-dessous. Les articles doivent être soumis sous forme de fichiers de traitement de texte, préférablement Word. Une version papier pourrait être requise pour les formules et graphiques.

- 1. **Présentation**
  - 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
  - 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
  - 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
  - 1.4 Les remerciements doivent paraître à la fin du texte.
  - 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.
- 2. **Résumé**

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

- 3. **Rédaction**
  - 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
  - 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme  $\exp()$  et  $\log()$  etc.
  - 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
  - 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
  - 3.5 Distinguer clairement les caractères ambigus (comme w,  $\omega$  ; o, O, 0 ; l, 1).
  - 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots.
- 4. **Figures et tableaux**
  - 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
  - 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).
- 5. **Bibliographie**
  - 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence. Exemple: Cochran (1977, page 164).
  - 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.
- 6. **Communications brèves**

Les documents soumis pour la section des communications brèves doivent avoir au plus 3 000 mots.

Volume 35, No. 4, December/décembre 2007

Bent JØRGENSEN & Peter X.K. SONG	Stationary state space models for longitudinal data.....	461
Petra BÜZKOVÁ & Thomas LUMLEY	Longitudinal data analysis for generalized linear models with follow-up dependent on outcome-related variables .....	485
Jae Kwang KIM & Jay J. KIM	Nonresponse weighting adjustment using estimated response probability.....	501
Jinhong YOU & Gemai CHEN	On inference for a semiparametric partially linear regression model with serially correlated errors.....	515
Alexander R. DE LEON & K.C. CARRIÈRE	General mixed-data model: extension of general location and grouped continuous models .....	533
Cinzia CAROTA	A family of power-divergence diagnostics for goodness-of-fit .....	549
Chiung-Yu HUANG, Jing QIN & Fei ZOU	Empirical likelihood-based inference for genetic mixture models .....	563
Xingwei TONG, Chao ZHU & Jianguo SUN	Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments.....	575
Jesse FREY	Distribution-free statistical intervals via ranked-set sampling.....	585
Heng LIAN	Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces.....	597
	Forthcoming papers/Articles à paraître.....	613
	Volume 36 (2008): Subscription rates/Frais d'abonnement .....	614



# JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

## Contents Volume 23, No. 4, 2007

Non-Bayesian Multiple Imputation

Jan F. Bjørnstad

433

Discussion

Ray Chambers

453

Ralf T. Münich

455

Chris Skinner

463

Seppo Laaksonen

467

Daniel Thorburn

477

Rejoinder

Jan F. Bjørnstad

485

The Interdependence of Determinants for the Strength and Direction of Social Desirability Bias in Racial

Attitude Surveys

Volker Stocké

493

The Perils of Interpreting Age Differences in Attitude Reports: Question Order Effects Decrease

with Age

Barbel Knäuper, Norbert Schwarz, Denise Park and Andreas Fritsch

515

Dependent Interviewing and Seam Effects in Work History Data

Annette Jackle and Peter Lynn

529

The Influence of End-Users on the Temporal Consistency of an International Statistical Process:

The Case of Tropical Forest Statistics

Alan Grainger

553

Book and Software Reviews

593

Editorial Collaborators

603

## Contents Volume 24, No. 1, 2008

The Morris Hansen Lecture 2004

Bridging the Gap: Moving to the 1997 Standards for Collecting Data on Race and Ethnicity

Jennifer H. Madans

1

Discussion

Clyde Tucker

13

Robert B. Hill

17

A Model-free Bayes Analysis of Stratification, Clustering and Regression in Finite Population Survey Data

Murray Aitkin

21

On the Covariance Between Related Horvitz-Thompson Estimators

John Wood

53

Probability Based Estimation Theory for Respondent Driven Sampling

Erk Volz and Douglas D. Heckathorn

79

Survey Experiences and Later Survey Attitudes, Intentions and Behaviour

Lars R. Bergman and Robert Brage

99

Temporal Disaggregation and Seasonal Adjustment

Tommaso Proietti and Filippo Moauro

115

New Approaches to Creating Data for Economic Geographers

Matthew Freedman, Julia Lane and Marc Roemer

133

Book and Software Reviews

157

All inquiries about submissions and subscriptions should be directed to [jos@scb.se](mailto:jos@scb.se)

### Remerciements

Montaquila, J., Bell, B., Mohadjer, L. et Rizzo, L. (1999). A methodology for sampling households late in a decade. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Montaquila, J., Mohadjer, L. et Khare, M. (1998). The enhanced sample design of the future National Health and Nutrition Examination Survey (NHANES). *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Les auteurs remercient le rédacteur associé et les examinateurs de leurs suggestions et commentaires constructifs qui leur ont permis d'améliorer considérablement l'article.

### Bibliographie

Bell, B., Mohadjer, L., Montaquila, J. et Rizzo, L. (1999). Creating a frame of newly constructed units for household surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Curtin, L.R., et Mohadjer, L. (2008). Design trade-offs for the National Health and Nutrition Examination Survey. *Proceedings of the Ninth Conference on Health Surveys Research Methods*, à paraître.

Wakseberg, J., et Mohadjer, L. (1991). Automation of within-household sampling. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Park, I., Dohmann, S., Montaquila, J., Mohadjer, L. et Curtin, L.R. (2006). Reducing the risk of data disclosure through area masking: Limiting biases in variance estimation. *Proceedings of the Survey Research Methods Section, American Statistical Association*.

individus à chaque emplacement de collecte pour essayer d'obtenir leur appui et l'utilisation des médias (journaux, télévision et radio) afin de rejoindre un nombre aussi grand que possible de PE. Il est, par conséquent, relativement facile de déterminer les comités compris dans l'échantillon de la NHANES. La composition raciale ou ethnique d'un comité, ainsi que la situation de région statistique métropolitaine ou non métropolitaine fournissent des renseignements suffisants pour apparier correctement une liste de comités connus à des groupes identifiés comme formant une grappe dans un comité dans le fichier de données à grande diffusion. Pour limiter la divulgation géographique, on recourt à des méthodes de permutation probabilistes des enregistrements au deuxième degré d'échantillonnage (permutation des segments) afin de créer des unités à variance masquée. L'objectif est de réduire le risque d'identifier des individus en masquant leur emplacement. Consulter Park, Dohmann, Montaquila, Mohadjer et Curtin (2006) pour une description des procédures de permutation appliquées à l'échantillon de la NHANES.

## 5. Sommaire et conclusion

Une caractéristique unique de la NHANES est l'examen médical complet effectué dans les CEM. En outre, l'enquête est conçue de manière à produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population des États-Unis, puisque la plupart des analyses des données de la NHANES sont faites pour des groupes d'âge particuliers, dans divers sous-groupes socioéconomiques de la population. Donc, le plan d'échantillonnage de la NHANES doit établir un équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail efficace pour le personnel d'examen du CEM, d'autre part, tout en maintenant les taux de réponse aussi élevés que possible. En outre, le plan doit être aussi rentable que possible, produire des échantillons annuels efficaces et permettre le cumul des échantillons au cours du temps pour les sous-domaines ou les maladies rares. De surcroît, le plan doit être souple pour permettre de modifier les paramètres clés, y compris les domaines d'échantillonnage, et les taux d'échantillonnage afin de répondre aux nouvelles questions en matière de santé.

Les exigences susmentionnées se traduisent par un plan d'échantillonnage très complexe dont certaines caractéristiques sont propres à la NHANES. En particulier, l'échantillon courant est conçu afin de produire des échantillons annuels et pluriannuels efficaces. La NHANES utilise des UPE pondérées et des mesures de taille de segment pour produire des échantillons autopondérés pour

chaque sous-domaine, tout en produisant une charge de travail efficace dans chaque UPE. Une fois que les tailles d'échantillon sont calculées, elles sont traitées comme des quotas. Les tailles d'échantillon sont strictement contrôlées dans chaque UPE afin que les opérations sur le terrain soient géographiques et efficaces. Un très grand échantillon de présélection est utilisé afin de surechantillonner la plupart des sous-domaines d'âge et de revenu, et le surechantillonnage des régions à forte concentration est utilisé pour certains sous-domaines minoritaires très rares. L'échantillon de PE est sélectionné selon une méthode pseudo-aléatoire afin de maximiser le nombre moyen de personnes sélectionnées par ménage, parce que ceci a semblé accroître le taux global de réponses lors des enquêtes précédentes.

Les défis décrits dans le présent article ont trait aux principaux aspects de la NHANES. Cette dernière possède de nombreux autres caractéristiques uniques dont il faut tenir compte lors de l'analyse des données. Par exemple, chaque échantillon annuel ne contient qu'un très petit nombre d'UPE, mais les données recueillies dans ces UPE ne sont pas aléatoirement d'une saison à l'autre. En particulier, s'il existe une interaction entre la saison et la région géographique pour une variable d'intérêt, le plan d'échantillonnage actuel de la NHANES ne permettra pas de l'estimer. Étant donné le petit nombre d'UPE dans chaque cycle de diffusion des données, tout couplage contextuel des données au niveau géographique doit se faire au centre de données de recherche du NCHS. Comme le nombre de sous-échantillons de la NHANES est élevé, il convient de prendre tout spécialement soin d'utiliser le poids de sous-échantillon approprié; ainsi, les estimations des cas de diabète non diagnostiqués doivent être calculées en utilisant le poids spécial pour le test effectué à jeun.

Afin de faciliter la bonne utilisation des CEM pour la collecte des données, aucun effort n'a été fait en vue de répartir aléatoirement l'échantillon d'UPE entre les périodes dans les échantillons annuels. Cependant, la dimension temporelle joue un rôle important dans certains indicateurs de la santé, telle que la nutrition. De surcroît, l'analyse des données sur la nutrition peut aussi être influencée par la nature complexe du plan d'échantillonnage et de la collecte des données. Des poids de sondage spécifiques construits pour les deux jours de collecte des données du rappel alimentaire d'examen selon le jour de la semaine. Un tutoriel sur Internet est développé à l'heure actuelle afin de faciliter l'analyse des données sur la nutrition de la NHANES. Un tutoriel général concernant l'analyse fondée sur le plan de sondage des données de la NHANES peut être obtenu à <http://www.cdc.gov/nchs/tutorials/>.



constructions neuves (c'est-à-dire des unités de logement construites depuis le recensement décennal le plus récent), le même taux d'échantillonnage produira un nombre trois fois plus grand d'interviews et d'examen médicaux que le nombre prévu au départ. Des changements aussi spectaculaires de la taille des segments n'est pas surprenante lorsque la période de collecte des données est ultérieure de plusieurs années au recensement décennal le plus récent pour lequel des fichiers de données sont disponibles.

Dans le cas de la NHANES, les tailles d'échantillons ne peuvent pas varier fortement, à cause des calendriers établis pour les CEM. Le sous-échantillonnage dans les UPE pour essayer d'obtenir des échantillons de même taille dans toutes les UPE n'est pas recommandé non plus, car il introduirait des facteurs de pondération inégaux qui réduiraient l'efficacité de l'échantillon.

Le programme de la NHANES a utilisé deux méthodes pour mettre à jour les mesures de taille des segments, à savoir 1) la création de segments de constructions neuves et 2) l'échantillonnage à deux phases pour mettre à jour la mesure de taille. Une troisième approche consistant à acheter des listes d'adresses commerciales pour mettre à jour la mesure de taille dans un plan d'échantillonnage à deux phases est à l'étude.

Sous l'approche des constructions neuves (Bell, Mohadjer, Montaquila et Rizzo 1999), les unités qui viennent d'être construites sont exclues des segments de région et de nouveaux segments sont créés en se basant sur l'information du U.S. Census Bureau sur les permis émis pour des constructions neuves depuis le recensement décennal le plus récent. Les segments de constructions neuves comprennent des grappes de permis de bâtir émis durant un ou plusieurs mois contigus par un bureau d'octroi de permis de bâtir. Les fichiers de la Building Permits Survey menée par le Census Bureau servent de sources de données sur le nombre de permis de bâtir résidentiels émis par les bureaux d'octroi de permis de bâtir.

L'échantillonnage à deux phases est utilisé dans un certain nombre d'applications statistiques. L'une d'elles est la mise à jour d'une base de sondage lorsque l'échantillon doit être sélectionné en fonction d'une mesure de taille, mais qu'une estimation fiable de la mesure de taille n'est pas disponible. Selon cette approche, un échantillon plus grand d'unités (de segments dans le cas de la NHANES) est sélectionné. Une valeur mise à jour de la mesure de taille est alors recueillie pour cet échantillon plus grand (également appelé échantillon de première phase). L'échantillon final d'unités (de segments) est sélectionné à partir de l'échantillon de première phase en utilisant la mesure de taille mise à jour.

À partir de 2000, la mesure de taille des segments de la NHANES a été mise à jour pour les emplacements de

collecte pour lesquels cette mise à jour semblait nécessaire) en utilisant une méthode d'échantillonnage à deux phases (Montaquila, Bell, Mohadjer et Rizzo 1999). Dans ces cas, les personnes chargées de dresser les listes de logements se rendent dans l'emplacement de collecte pour obtenir un dénombrement des unités de logement (UL) dans chaque segment de l'échantillon de première phase. Parant de ces dénombrements, une mesure de taille mise à jour reflétant le ratio du nombre actuel d'UL au nombre prévu d'UL est calculée pour chaque segment de première phase. L'échantillon final de segments est alors sélectionné par sous-échantillonnage des segments de première phase en utilisant la mesure de taille mise à jour.

*Risque de divulgation des données par identification géographique* - De nos jours, les questions de confidentialité et le risque de divulgation des données posent de réels défis aux organismes qui parviennent les enquêtes. La capacité d'identifier les répondants à une enquête, d'après des combinaisons uniques de variables disponibles dans un seul fichier de données ou sur des coupplages de divers bases de données, est un grave sujet de préoccupation. Il en est particulièrement ainsi de la NHANES, étant donné la grande quantité de renseignements de nature délicate recueillis sur chaque personne échantillonnée et le petit nombre d'UPE dans l'échantillon. Par conséquent, le risque de divulgation géographique et la divulgation d'après les caractéristiques individuelles. Diverses méthodes (diffusion limitée ou suppression de données) sont utilisées par la NHANES pour masquer les caractéristiques individuelles posant un grand risque de permettre d'identifier des personnes qui font partie de l'échantillon. Les éléments de données délicats, à diffusion limitée ou non diffusés, sont consultables dans un centre de données de recherche. À l'heure actuelle, seules les estimations nationales peuvent être produites d'après les fichiers de microdonnées à grande diffusion, et les analyses géographiques détaillées doivent être faites dans le Centre de données de recherche.

Bien que l'on ne puisse produire que des estimations nationales, l'estimation directe des erreurs d'échantillonnage pour ces estimations nécessite la diffusion des variables de plan, comme les identificateurs de strate et d'UPE. Habituellement, ces variables indiquent qu'un groupe de personnes échantillonnées vivent toutes dans le même comté, mais n'identifie pas le comté. La divulgation géographique est une cause de souci particulière, car, dans le cas de la NHANES 1) le nombre d'UPE est petit, 2) les UPE sont limitées géographiquement à un comté et 3) de nombreuses démarches de relations communautaires sont menées dans chaque UPE en vue d'améliorer les taux de réponse. Le programme de relations communautaires et comprend la prise de contact avec divers organismes et





Les paragraphes qui suivent décrivent brièvement les caractéristiques uniques du plan d'échantillonnage de la NHANES.

#### *Mise en grappes et mesures de taille* - Dans la NHANES,

la taille d'échantillon doit être suffisamment grande pour produire une charge de travail efficace dans chaque UPE, compte tenu du temps et du coût de déplacement d'un CEM entre deux emplacements de collecte et du temps nécessaire pour monter le CEM et le démonter pour le déplacement. Selon l'expérience acquise lors de réalisations antérieures de la NHANES, un nombre moyen de 340 PE examinées est un nombre approximativement optimal qui produit le nombre maximal d'UPE, tout en maintenant la taille d'échantillon dans chaque région suffisamment grande pour justifier les coûts du déménagement du CEM. En outre, dans le cas de la NHANES, les UPE sont habituellement définies comme des comtés individuels afin de réduire le temps de déplacement des répondants pour se rendre au CEM et donc accroître la probabilité d'obtenir des taux de réponse élevés.

L'échantillon de la NHANES est conçu pour produire un échantillon autopondéré pour chaque sous-domaine échantillonné, tout en créant une charge de travail efficace dans chaque UPE. Les UPE et les segments sont sélectionnés avec une probabilité proportionnelle à une mesure de taille pondérée qui reflète la population de l'UPE dans les sous-domaines d'intérêt. La probabilité de sélection d'une UPE détermine le taux maximal auquel les personnes résidant dans cette UPE particulière peuvent être sélectionnées. Voir le document *Vital and Health Statistics, Series 2, No. 113, September 1992, CDC/NCHS*, consultable à <http://www.cdc.gov/nchs/products/pubs/pubd/series/str02/120-101/120-101.htm> pour une description des mesures de taille utilisées dans le cadre de la NHANES.

*Echantillons annuels et stratification* - Un moyen de réaliser des échantillons annuels représentatifs de la population nationale consiste à sélectionner un échantillon indépendant d'UPE. Étant donné le nombre limité d'UPE de la NHANES et le fait que ces dernières sont sélectionnées avec probabilité proportionnelle à la taille, cette approche donnerait vraisemblablement lieu à un chevauchement important des UPE d'une année à l'autre. Le chevauchement des échantillons, même au niveau des UPE, pourrait entraîner une perte de précision des estimations calculées d'après les données de l'enquête si les échantillons de plusieurs années de référence sont combinés (à cause de l'accroissement de la mise en grappes dans l'échantillon). Donc, au lieu d'échantillonner les UPE indépendamment chaque année, il a été décidé de sélectionner pour la NHANES un échantillon de six ans, à partir d'une structure hiérarchique de strates principales et secondaires (décrite plus loin), puis à affecter une UPE de chacune des

strates principales à chaque année. Cette structure hiérarchique de l'échantillon de six ans évite le chevauchement des UPE non autoreprésentatives durant les six années. L'échantillon de six ans de la NHANES est sélectionné selon un plan stratifié à deux UPE par strate qui a été élaboré en ayant pour objectif principal l'efficacité de l'échantillon de six ans, ainsi que celle des échantillons plurianuels. Le plan de stratification est conçu de façon que les UPE formant les échantillons annuels et plurianuels soient réparties uniformément en fonction de certaines caractéristiques géographiques et démographiques.

Le plan d'échantillonnage de la NHANES comprend (jusqu'à 2006 inclusivement) 18 UPE autoreprésentatives. Ces UPE variaient de celles qui étaient autoreprésentatives pour les échantillons annuels à celles qui l'étaient pour les échantillons de trois ans ou de six ans. Ces UPE ont été affectées de telle façon que le nombre d'UPE autoreprésentatives des échantillons de trois ans étant espacées de sentatives des échantillons de trois ans étant espacées de trois ans. Chaque strate principale comprenait six strates secondaires et une UPE a été sélectionnée dans chacune de ces strates finales. Dans chaque strate principale, les strates secondaires ont été appariées pour créer des pseudo-strates. Chaque paire a été assignée aléatoirement à l'étude avec un intervalle de trois ans. L'affectation des paires aux ensembles particuliers d'années de référence et l'affectation des années de référence dans les paires a été faite aléatoirement dans la première strate principale et le même schéma a été suivi dans toutes les autres.

Ce plan de stratification a donné un échantillon de 72 UPE non autoreprésentatives qui permettent de produire des estimations annuelles et plurianuelles efficaces sans compromettre l'efficacité des estimations sur six ans. L'échantillon de six ans est obtenu selon un plan d'échantillonnage d'une UPE par strate secondaire (ou un plan d'échantillonnage de deux UPE par pseudo-strate) et chaque échantillon annuel, selon un plan d'échantillonnage d'une UPE par strate principale. En plus, ce plan offre la souplesse nécessaire pour répondre aux changements d'exigences d'échantillonnage (si un nouvel échantillon doit être sélectionné), puisque pendant les trois premières années, l'échantillon correspond à un plan d'échantillonnage d'une UPE par pseudo-strate.

*Nombre maximisé de personnes échantillonnées par ménage* - Après avoir obtenu l'échantillon de ménages présélectionnés, on sélectionne un échantillon de personnes qui seront interviewées et soumises à l'examen médical. La liste de tous les membres admissibles d'un ménage est dressée et un sous-échantillon de personnes est sélectionné en fonction du sexe, de l'âge, de la race ou de l'ethnicité et du revenu (toutes les femmes enceintes sont sélectionnées avec certitude). Les PE sont sélectionnées à des taux établis





*Charge de travail des centres d'examen mobiles (CEM) - Les CEM sont constitués de quatre remorques spécialement conçues et équipées et contiennent tout l'équipement médical. Chaque remorque mesure environ 45 pieds de long et 10 pieds de large. Un camion tracteur détachable conduit les remorques d'un emplacement à un autre. Les CEM se rendent aux divers emplacements de collecte à travers le pays. Les remorques sont installées côte à côte et jointes par des passerelles fermées. L'espace à l'intérieur du CEM est divisé en salles pour permettre le respect de la vie privée durant les examens et les interviews. L'examen comprend diverses évaluations et mesures physiques et dentaires, des analyses de laboratoire et des interviews sur la santé.*

Étant donné les difficultés logistiques associées à l'utilisation des CEM, pour chaque emplacement échantillonné, la taille d'échantillon doit être déterminée d'avance et considérée comme fixe, afin que la planification des opérations sur le terrain soit efficace et pratique. En outre, il est nécessaire d'établir un calendrier ferme pour chaque emplacement de collecte, afin que les rendez-vous puissent être pris pour les examens. Il est impossible de modifier le calendrier, car celui-ci doit être coordonné avec les visites du CEM à d'autres emplacements dont le calendrier est également préétabli.

*Taux de réponse -* L'obtention de taux de réponse élevés est une préoccupation dans le cas de presque toutes les enquêtes par sondage. Dans celui de la NHANES, le défi est particulièrement grand, étant donné la portée des interviews et des examens. L'offre d'une rémunération a été utilisée comme moyen d'améliorer les taux de réponse. En outre, la NHANES comprend un programme de relations communautaires important englobant des contacts avec les organisations locales et les personnes dont il faut obtenir la coopération, ainsi que la couverture dans les médias locaux afin de joindre un aussi grand nombre que possible de PE. Dans le contexte des questions soulevées par le plan d'échantillonnage, une approche qui s'est avérée avoir un effet favorable sur les taux de réponse a été la sélection d'échantillons de plus grande taille dans les ménages échantillonnés. L'un des facteurs soupçonnés d'être à l'origine de l'accroissement des taux de réponse dans les ménages comptant plusieurs PE est que chaque personne reçoit un dédommagement pour son temps et sa participation et qu'il est généralement plus commode pour les membres du ménage de se rendre ensemble au CEM. Le tableau 1 donne le taux de réponses à la composante de l'examen des PE provenant des ménages dans lesquels une seule personne a été sélectionnée comparativement au taux pour les PE provenant des ménages où plusieurs personnes ont été sélectionnées. Comme le montre le tableau, les taux de réponse augmentent d'environ 4 à 7 % selon le type de ménage.

Par conséquent, la NHANES est conçue en vue de maximiser le nombre de PE par ménage. Cette approche est faisable dans le cas d'études de ce genre où l'échantillon est constitué d'un grand nombre de sous-domaines. Autrement dit, l'effet de la mise en grappes dans les ménages n'est pas très préoccupant, parce que la plupart des analyses sont faites dans des sous-domaines âge-sexe-particuliers (ou dans certains groupes limités de sous-domaines) et que la mise en grappes dans les ménages est généralement faible au niveau du sous-domaine. Le nombre moyen de PE sélectionnées par ménage (dans les ménages où au moins une PE a été sélectionnée) dans les domaines définis d'échantillonnage varie de 1 à 1,24 dans l'échantillon pour 1999 à 2006. Le regroupement des domaines, afin de réduire leur nombre entre 12 et 15, en fonction de l'âge et/ou de la race/ethnicité donne des nombres moyens variant de 1,01 à 1,37 PE par ménage. Par conséquent, un certain niveau de mise en grappes existe dans la mesure où des domaines groupés sont utilisés pour l'analyse. Il convient de souligner que l'échantillon de PE est utilisé fondamentalement pour la analyse au niveau des PE (par exemple, statistiques sur la santé et la nutrition). La mise en grappes des PE est évidemment plus importante au niveau de la famille et du ménage. Cependant, des variables au niveau du ménage ou de la famille sont utilisées pour ce genre d'analyse (par exemple, niveau de poussière dans le ménage, revenu familial ou assurance familiale). Consulter Curtin et Mohadjer (2008) pour une discussion de l'effet de la mise en grappes et des probabilités inégales de sélection des sous-domaines sur les niveaux de précision des diverses estimations.

**Tableau 1**  
Examen des taux de réponse selon le nombre de PE dans le ménage, par type de ménage, dans l'échantillon de la NHANES pour 1999 à 2006

Type de ménage	Nombre de PE sélectionnés par ménage	Taux de réponse (%)	Deux PE	Une PE ou plus	Deux PE ou plus
Noir/mexicain	4 892	20 222	76,5	82,3	84,5
Autre à faible revenu <sup>1</sup>	1 362	3 349	77,6	84,5	84,5
Autre pas à faible revenu	5 597	15 508	68,8	72,6	72,6
Le groupe Autre comprend toutes les PE qui ne sont ni noires ni mexicaines. Le seuil de faible revenu est fixé à 130 % du seuil de pauvreté.					

*Ratio des coûts -* Dans les échantillons d'enquête aréolaire, le coût des opérations de collecte des données sur le terrain comprend le coût d'établissement des listes d'unités de logements, la présélection des ménages afin de repérer les répondants admissibles et la réalisation de l'interview pour recueillir les données. Dans le cas de la NHANES, la phase de collecte des données comprend



## 2. Principaux objectifs de la NHANES

La NHANES est une enquête annuelle permanente sur la santé réalisée auprès de la population civile ne résidant pas en établissement des États-Unis. Ses principaux objectifs sont : 1) estimer la prévalence nationale de certaines maladies et certains facteurs de risque, 2) estimer les distributions de référence dans la population nationale de certains paramètres de la santé et contaminants présents dans l'environnement, 3) décrire et étudier les raisons des tendances séculaires de certaines maladies et certains facteurs de risque, 4) contribuer à la compréhension des causes des maladies, 5) étudier l'évolution naturelle de certaines maladies, 6) étudier la relation entre le régime alimentaire, la nutrition, l'environnement, la génétique et la santé et 7) explorer les questions de santé publique naissantes.

### 3. Principaux facteurs ayant une incidence sur le plan d'échantillonnage

Comme nous l'avons mentionné plus haut, une caractéristique unique de la NHANES est qu'un examen médical complet est effectué dans les centres d'examen mobiles. En outre, le plan doit produire des tailles d'échantillon efficaces pour un grand nombre de sous-domaines de la population générale. Beaucoup de caractéristiques de la santé et de la nutrition diffèrent considérablement selon l'âge, le sexe et la race ou ethnité, et dépendent aussi de la situation de revenu. Par conséquent, la plupart des analyses des données de la NHANES sont effectuées pour des groupes d'âge particuliers dans divers sous-groupes socio-économiques de la population. L'enquête est donc conçue afin de produire des tailles d'échantillon efficaces pour un très grand nombre de sous-domaines de la population des États-Unis.

En général, le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaine efficaces, d'une part, et d'une charge de travail pouvant être gérée par le personnel du CEM, d'autre part, tout en maintenant les taux de réponse aussi élevés que possible. Plus précisément, le plan d'échantillonnage de la NHANES vise à 1) obtenir des tailles présélectionnées d'échantillons autopondérés pour un ensemble d'environ 75 sous-domaines prédéfinis, 2) produire des tailles d'échantillon par UPE dominant lieu à une charge de travail gérable pour les intervieweurs et le personnel du CEM, 3) obtenir des échantillons susceptibles de produire des taux de réponse élevés, 4) être aussi rentable que possible, 5) produire des échantillons annuels efficaces, 6) permettre le cumul d'échantillons au cours du temps, surtout pour les sous-domaines ou les maladies rares et

La suite de la section est consacrée à un bref résumé de l'incidence de chacun de ces sept objectifs sur la conception et la mise en œuvre de la NHANES.

*Sous-domaines de la NHANES* - Le plan d'échantillonnage de la NHANES permet d'atteindre un niveau présélectionné de précision pour les données transversales et les comparaisons au cours du temps pour un ensemble de sous-domaines prédéfinis. Plus précisément, 77 domaines d'échantillonnage (dans l'échantillon de 2006) sont définis en fonction de la race/ethnité, le sexe, l'âge, le revenu et la situation de grossesse. Les Noirs, les Mexicains, les enfants très jeunes, les adolescents, les personnes âgées, les femmes enceintes et les personnes à faible revenu sont suréchantillonnés.

Quand les estimations de totaux d'univers pour l'ensemble de la population sont considérées comme étant de la plus haute importance, la meilleure estimation disponible du total de population est utilisée comme mesure de taille dans le processus d'échantillonnage. Dans le cas de la NHANES, où l'on s'intéresse à des sous-domaines de la population totale, une autre mesure de taille est nécessaire pour améliorer l'exacitude des estimations et permettre de mieux contrôler la taille de l'échantillon. La section 4 décrit la mesure de taille utilisée pour l'échantillonnage des UPE et des segments dans le cadre de la NHANES.

L'objectif du suréchantillonnage (en utilisant des probabilités de sélection différentes) est de tirer un échantillon contenant des nombres proportionnellement plus élevés de membres de certains sous-domaines de population que n'en contient la population. Le but est d'obtenir des tailles d'échantillon adéquates pour faire des inférences pour des sous-domaines représentant une proportion relativement faible de l'univers d'intérêt total et de le faire de façon à réduire au minimum les variances, compte tenu du budget de l'enquête. Diverses stratégies de suréchantillonnage sont utilisées, selon le domaine d'intérêt. Par exemple, le suréchantillonnage des sous-populations minoritaires est réalisé par stratification des régions géographiques et par sélection de ces groupes minoritaires et par sélection des segments à un taux plus élevé dans les régions à forte concentration. Par ailleurs, un grand échantillon de sélection peut être nécessaire pour le suréchantillonnage des personnes appartenant à des groupes d'âge particuliers. La sous-section sur les ratios des coûts qui figure plus bas décrit pourquoi les procédures de suréchantillonnage appliquées dans le cas de la NHANES diffèrent de celles habituellement utilisées pour de nombreuses enquêtes par sondage à base aréolaire.



en établissement ne font pas partie de la population d'intercense. L'échantillonnage de la NHANES ne se fait pas selon un plan à probabilité égale; les fractions d'échantillonnage sont établies de manière à « suréchantillonner » les Mexicano-Américains (et les Hispaniques dans les échantillons de 2007 et au-delà), les Américains de race noire, les Américains de race blanche et autres à faible revenu, les personnes de moins de 20 ans et les personnes de plus de 60 ans. Un plan d'échantillonnage à quatre degrés est utilisé. Les unités primaires d'échantillonnage (UPE), souvent appelées emplacements de collecte (*stands*), sont sélectionnées à partir d'une base de sondage comprenant tous les comtés des États-Unis. Les UPE sont principalement des comtés uniques; dans quelques cas, des comtés adjacents sont fusionnés afin que la taille des UPE demeure supérieure à une taille minimale fixée. La base de sondage de la NHANES contient près de 3 000 UPE qui sont échantillonnées avec probabilité proportionnelle à une mesure de taille (PPT). Chaque échantillon annuel comprend 15 emplacements de collecte.

Le deuxième degré d'échantillonnage correspond à la sélection de segments de régions (*area segments*) constitués d'îlots de recensement ou de combinaisons d'îlots. Comme les UPE ne sont pas toutes de même taille, la taille des segments et le nombre de segments par UPE présentent une certaine variabilité. Chaque segment est formé d'environ 150 ménages (ou unités de logements) en moyenne, environ 5 000 segments sont créés dans chaque UPE et 24 segments sont échantillonnés en moyenne. L'échantillonnage est conçu de façon à ce que la taille d'échantillon soit approximativement la même dans chaque UPE et la plupart des UPE comptent exactement 24 segments. Les segments sont également sélectionnés avec probabilité proportionnelle à la taille. Les mesures de taille des segments, lorsqu'elles sont combinées aux taux de sous-échantillonnage dans les segments, fournissent des nombres approximativement égaux de personnes échantillonnées (PE) par segment, quoique la variation relative de la charge de travail soit plus grande dans les segments que dans les UPE.

Le troisième degré d'échantillonnage consiste à sélectionner les ménages et les logements collectifs non institutionnels, comme les dortoirs. Dans une UPE donnée, après la sélection de segments, toutes les unités de logement (UL) comprises dans les segments échantillonnés sont répertoriées et un sous-échantillon de ménages et de logements collectifs compris dans les UL sont désignés pour une présélection afin d'identifier les PE éventuelles pour l'interview et l'examen médical. Les PE comprises dans les ménages ou les logements collectifs constituent le quatrième degré d'échantillonnage. Tous les membres admissibles d'un ménage sont répertoriés et un sous-échantillon de personnes est sélectionné. Les taux de sous-échantillonnage des

ménages dans les segments et de personnes dans les ménages sont prédéterminés. La combinaison de la présélection et des taux d'échantillonnage différentiels fournit l'accroissement de taille d'échantillon pour les sous-domaines démographiques d'intérêt (âge, sexe, race/ethnicité et revenu). Ainsi, dans les 30 UPE dans lesquelles les données ont été recueillies durant le cycle de collecte de deux ans de 2005-2006, 716 segments ont été sélectionnés et 26 529 ménages ont été tirés pour la présélection. Après la présélection pour déterminer la composition selon l'âge, le sexe et la race/ethnicité et la situation de faible revenu, une ou plusieurs personnes ont été sélectionnées dans 6 372 ménages pour faire partie de l'échantillon. En tout, 12 862 personnes ont été sélectionnées, et parmi celles-ci, 9 950 ont répondu à l'interview et subi les examens.

Les examens de la NHANES requièrent du personnel hautement spécialisé, ainsi que l'analyse en laboratoire des échantillons prélevés. Par conséquent, la mise en œuvre des composantes de l'examen médical peut être très coûteuse. Afin de limiter les coûts et de réduire le fardeau de réponse, certaines composantes de l'examen médical ne sont administrées qu'à un sous-échantillon des répondants qui se présentent au CEM. Un seul algorithme de sous-échantillonnage sert à contrôler le degré de chevauchement entre les divers sous-échantillons afin qu'il soit possible d'analyser les corrélations entre les divers examens et composantes de laboratoire. L'affectation des PE aux sous-échantillons est déterminée entièrement avant qu'elles se

présentent au CEM.

Les données recueillies dans le cadre des enquêtes de la NHANES ont joué un rôle extrêmement important dans l'obtention des renseignements nécessaires sur la santé et l'état nutritionnel de la population des États-Unis. Par conséquent, à partir de 1999, la NHANES est devenue une enquête annuelle permanente (Mohadjer et Khare 1998). Il est essentiel d'accorder beaucoup d'attention à l'établissement et à la tenue à jour d'un plan d'échantillonnage efficace dans le cas d'une enquête aussi importante et complexe. Le présent article décrit les défis posés par la conception et la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les multiples objectifs de la NHANES. Il porte sur le plan d'échantillonnage utilisé jusqu'en 2006 (afin de répondre aux nouvelles exigences analytiques, certains aspects du plan ont été modifiés à partir de 2007). La section 2 décrit les principaux objectifs de l'enquête et la section 3 donne un aperçu des facteurs les plus importants ayant une incidence sur le plan d'échantillonnage. La section 4 décrit les caractéristiques uniques du plan d'échantillonnage de la NHANES. Enfin, la section 5 résume brièvement l'article.

# Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey

Leyla Mohadjer et Lester R. Curtin

## Résumé

La National Health and Nutrition Examination Survey (NHANES) des États-Unis, une série de programmes ayant trait à la santé parrainés par le National Center for Health Statistics des États-Unis. Une caractéristique unique de la NHANES est que tous les répondants de l'échantillon sont soumis à un examen médical complet. Afin de normaliser la façon dont ils sont effectués, ces examens ont lieu dans des centres d'examen mobiles (CEM). L'examen comprend des mesures physiques, des tests tels que l'examen de la vue et des dents, ainsi que le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. La NHANES est une enquête annuelle continue sur la santé réalisée auprès de la population civile des États-Unis ne résidant pas en établissement. Les principaux objectifs analytiques comprennent l'estimation du nombre et du pourcentage de personnes dans la population des États-Unis et dans des sous-groupes désignés qui présentent certaines maladies et certains facteurs de risque. Le plan d'échantillonnage de la NHANES doit permettre d'établir un juste équilibre entre les exigences liées à l'obtention d'échantillons annuels et pluriannuels efficaces et la souplesse requise pour pouvoir modifier les paramètres essentiels du plan afin de mieux adapter l'enquête au besoin des chercheurs et des décideurs qui élaborent les politiques en matière de santé. Le présent article décrit les défis associés à la conception et à la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les objectifs de la NHANES.

Mots clés : Échantillonnage à plusieurs degrés; échantillonnage par domaine; mesure pondérée de la taille; centres d'examen mobiles.

## 1. Introduction

La National Health and Nutrition Examination Survey (NHANES) fait partie d'une série de programmes liés à la santé parrainés par les Centers for Disease Control and Prevention des États-Unis par l'entremise du National Center for Health Statistics (NCHS). La NHANES est utilisée depuis plus de 45 ans pour évaluer l'état de santé et l'état nutritionnel de la population civile des États-Unis ne résidant pas en établissement. Les données recueillies servent à estimer la prévalence des principales maladies et des principaux facteurs de risque de maladie. Les données sur la nutrition permettent une surveillance temporelle de la population nationale en ce qui a trait à des facteurs tels que le régime alimentaire, le taux de cholestérol, l'hypertension, la carence en fer, l'anémie et l'obésité. La NHANES est également conçue en vue d'évaluer la relation entre le régime alimentaire, la santé et l'environnement, afin de pouvoir établir le lien entre les évaluations nutritionnelles et des affections comme la maladie cardiovasculaire, le diabète, l'hypertension et l'ostéoporose.

La collecte des données de la NHANES comprend au moins trois étapes, à savoir un questionnaire de sélection des ménages, une interview et un examen médical. L'objectif principal du questionnaire de sélection est de déterminer si un membre du ménage est admissible à l'interview et à l'examen médical. Le questionnaire de sélection vise à recueillir des renseignements de base sur la composition et

les caractéristiques démographiques du ménage. L'interview est conçue pour recueillir des données au niveau du ménage, de la famille et de la personne sur les caractéristiques démographiques et socioéconomiques, la santé et les caractéristiques nutritionnelles. À la fin de l'interview, il est demandé au répondant de participer à un examen médical. Afin de normaliser la façon dont ils sont administrés et les protocoles, ces examens se déroulent dans un centre d'examen mobile (CEM) spécialement conçu et équipé. L'examen comprend des mesures physiques, des tests comme un examen des yeux et des dents, des mesures physiologiques et le prélèvement d'échantillons de sang et d'urine pour des analyses biologiques. Le site Web de la NHANES (<http://www.cdc.gov/nchs/nhanes.htm>) fournit des renseignements détaillés sur les composantes médicales de l'enquête.

L'élaboration d'un plan d'échantillonnage efficace a nécessité la résolution de plusieurs questions de conception particulières à la NHANES en plus de celles qui se posent habituellement en échantillonnage. Le présent article traite des aspects uniques et compliqués du plan d'échantillonnage de la NHANES. Néanmoins, nous estimons qu'il est utile de commencer par un résumé général de ce plan d'échantillonnage, ce que nous faisons ci-après, avant de discuter de ses caractéristiques uniques.

L'échantillon de la NHANES représente l'ensemble de la population civile ne résidant pas en établissement des États-Unis. Les militaires d'active et les personnes placées

- Tetl, M.F., Roe, B. et Vayda, M. (2005). Incentive effects on response rates, data quality, and survey administration costs. *International Journal of Public Opinion Research*, 18, 364-373.
- Tzamourani, P., et Lynn, P. (1999). The Effect of Monetary Incentives on Data Quality - Results from the British Social Attitudes Survey 1998 Experiment. CREST document de travail No. 73. Oxford : University of Oxford.
- Ward, R.K., Bogges, S., Selvaev, K. et McMahon, M.F. (2001). The use of targeted incentives to reluctant respondents on response rates and data quality. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA : American Statistical Association.
- Yu, J., et Cooper, H. (1983). A quantitative review of research design effects on response rates to questionnaires. *Journal of Marketing Research*, 20, 36-44.
- Williamack, D.K., Schuman, H., Pennell, B.-E. et Lepkowski, J.M. (1995). Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly*, 59, 1 mars 1995, 78-92.



## Remerciements

Nous remercions Iain Noble, du Department for Children Schools and Families, d'avoir facilité l'accès aux données et formulé des commentaires au sujet d'une version antérieure de l'article, Tim Thair et Rory Fitzgerald, de leurs réponses patientes et constructives à nos questions sur les données, Noah Uhlig et Heather Laurie de leurs commentaires et Mark Bryan de ses conseils pour le calcul des interactions dans les modèles non linéaires. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du Department for Education and Skills.

## Bibliographie

- Remerciements**
- Nous remercions Iain Noble, du Département for Children Schools and Families, d'avoir facilité l'accès aux données et formulé des commentaires au sujet d'une version antérieure de l'article, Tim Thair et Rory Fitzgerald, de leurs réponses patientes et constructives à nos questions sur les données, Noah Unthig et Heather Laurie de leurs commentaires et Mark Bryan de ses conseils pour le calcul des interactions dans les modèles non linéaires. Les opinions exprimées sont celles des auteurs et ne représentent pas forcément celles du Département for Education and Skills.
- Bibliographie**
- AAPOR (2006). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 4<sup>ème</sup> Edition : American Association for Public Opinion Research.
- Armstrong, J.S. (1975). Monetary incentives in mail surveys. *Public Opinion Quarterly*, 39, 111-116.
- Berk, M.L., Mahbowetz, N.A., Ward, E.P. et White, A.A. (1987). The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics*, 3, 449-457.
- Church, A.H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, 57, 62-79.
- Davern, M., Rockwood, T.H., Sherrod, R. et Campbell, S. (2003). Prepaid monetary incentives and data quality in face-to-face interviews: Data from the 1996 survey of income and program participation incentive experiment. *Public Opinion Quarterly*, 67, 139-147.
- Downs, B. (1999). *Incentive Use in Panel Surveys*. Internal Census Bureau Memorandum. Washington, DC : US Census Bureau.
- Fox, R.J., Crask, M.R. et Kim, J. (1988). Mail survey response rate: A meta-analysis of selected techniques for inducing response. *Public Opinion Quarterly*, 52, 467-491.
- Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25, 161-178.
- Goyder, J. (1994). An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society*, 36, 360-366.
- Groves, R., Singer, E. et Comins, A. (2000). Leverage-salience theory of survey participation: Description and an illustration. *Public Opinion Quarterly*, 64, 299-308.
- Groves, R.M., et Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.
- Hopkins, K.D., et Gulickson, A.R. (1992). Response rates in survey research: A meta-analysis of the effects of monetary gratuities. *Journal of Experimental Education*, 61, 52-62.
- James, J.M., et Bolstein, R. (1990). The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly*, 54, 1 septembre 1990, 346-361.
- Kennedy, P. (2003). *A Guide to Econometrics*. Oxford : Blackwell.
- Laurie, H. (2007). The effect of increasing financial incentives in a panel survey: An experiment on the british household panel survey. Wave 14. *ISER Working Paper No. 2007-05*. Colchester : University of Essex.
- Laurie, H., et Lynn, P. (sous presse). The use of respondent incentives on longitudinal surveys. Dans *Methodology of Longitudinal Surveys*, (Ed. P. Lynn), Chichester : Wiley.
- Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA : Sage.
- Lynn, P., Pardon, S., Hedges, B. et McAleese, I. (1994). England and Wales Youth Cohort Study: An Assessment of Alternative Weighting Strategies. Employment Department Research Series YCS No. 30. Sheffield : Employment Department.
- Mack, S., Huggins, V., Keahley, D. et Sundukchi, M. (1998). Do monetary incentives improve response rates in the survey of income and programme participation? *Proceedings of the Survey Research Methods Section*. American Statistical Association.
- Alexandria, VA : American Statistical Association.
- Martin, E., Abreu, D. et Winters, F. (2001). Money and motive: Effects of incentives on panel attrition in the survey of income and program participation. *Journal of Official Statistics*, 17, 267-284.
- Mason, R., Lesser, V. et Traugott, M.W. (2002). Effect of item nonresponse on nonresponse error and inference. Dans *Survey Nonresponse*, (Eds. R.M. Groves, D.A. Dillman, J.T. Eltinge, et R.J.A. Little). New York : John Wiley & Sons, Inc., 149-161.
- Norton, E.C., Wang, H. et Ai, C. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4, 154-167.
- Rodgers, W. (2002). Size of incentive effects in a longitudinal study. *Proceedings of the Survey Research Methods Section*. American Statistical Association. Alexandria, VA : American Statistical Association.
- Russell, N., et Phelps, A. (2001). Youth Cohort Study Cohort 10 Sweep 1 : Rapport technique. Londres : Taylor Nelson Sofres.
- Ryu, E., Couper, M.P., et Marans, R.W. (2006). Survey incentives: Cash vs. in-kind, face-to-face vs. mail: Response rate vs. nonresponse error. *International Journal of Public Opinion Research*, 18, 89-106.
- Shettle, C., et Mooney, G. (1999). Monetary incentives in U.S. government surveys. *Journal of Official Statistics*, 15, 231-250.
- Singer, E., Hoewyk, J.V., Gebler, N., Raghunathan, T. et McGonagle, K. (1999). The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics*, 15, 217-230.
- Singer, E., Van Hoewyk, J. et Maher, M.P. (1998). Does the payment of incentives create expectation effects? *Public Opinion Quarterly*, 62, 1 juin 1998, 152-164.
- Singer, E., Van Hoewyk, J. et Maher, M.P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly*, 64, 1 juin 2000, 171-188.
- Stratford, N., Simmonds, N. et Nicolaas, G. (2003). *National Travel Survey 2002: Report on Incentives Experiment*. Londres : National Centre for Social Research.
- Statistique Canada, N° 12-001-X à un catalogue

## 7. Résumé et discussion

une prime inconditionnelle. Les attentes formées sur la base des traitements incitatifs antérieurs pourraient fort bien signifier que les changements ont un effet durable, si ceux-ci réduisent la valeur de la prime d'encouragement aux yeux des primes d'encouragement n'ont pas eu d'effet sur le du répondant (voir, Singer, Van Hoewyk et Maier 1998).

Les primes d'encouragement ont eu un effet sur le biais d'attrition. Nous n'avons toutefois pas pu évaluer l'effet sur le biais de non-réponse à la vague 1. Idéalement, nous évaluerions à la fois la grandeur du biais dû à la non-réponse à la vague 1 et dû à l'attrition subséquente, et les effets des primes d'encouragement sur les deux. Il est possible que la non-réponse à la vague 1 soit plus nuisible en ce qui concerne le biais que l'attrition ultérieure, surtout dans les études telles que celle décrite ici, dont le taux de réponse initial est faible. Le cas échéant, l'effet des primes d'encouragement sur le biais à la vague 1 pourrait être plus important que tout effet sur le biais d'attrition. En outre, la discussion des effets des primes d'encouragement sur le biais d'attrition s'est concentrée entièrement sur les caractéristiques observées et, si les primes ne présentaient pas d'effet différentiel en ce qui concerne ces caractéristiques, elles pourraient néanmoins avoir des effets différents en ce qui concerne les facteurs observés. Le cas échéant, l'offre de primes d'encouragement aux répondants pourrait introduire un biais de sélection d'échantillon dans les estimations multivariées, si les caractéristiques inobservées déterminant l'absence de sensibilité aux primes d'encouragement sont corrélées aux résultats mesurés par l'enquête (Kennedy 2003). Par exemple, si la sensibilité aux primes d'encouragement dépend de préférences temporelles pour l'argent et que ce facteur détermine aussi la décision d'arrêter d'étudier et de travailler à la place, alors les modèles des déterminants des résultats en matière d'éducation produiront des estimations biaisées.

Enfin, il se dégage peu de preuves que l'échantillon de répondants est devenu moins sensible aux primes d'encouragement d'une vague à l'autre, à mesure que des membres de l'échantillon éventuellement moins déterminés à participer ont abandonné l'enquête. Ce constat concorde avec celui de Laurie (2007), qui rapporte qu'un accroissement de la valeur d'une prime d'encouragement dans la British Household Panel Survey a augmenté significativement la réponse annuelle, de l'ordre de 95 %, était déjà élevé. Puisque des études antérieures ont montré que les effets d'une prime d'encouragement unique peuvent se transmettent la réponse, même après 14 vagues du panel, dont le taux de réponse annuel, de l'ordre de 95 %, était déjà élevé. Puisque des études antérieures ont montré que les effets d'une prime d'encouragement unique peuvent se transmettre d'une vague à l'autre (James 1997; Laurie et Lynn sous presse; Mack et coll. 1998), un test formel des effets marginaux des primes d'encouragement nécessiterait toutefois des comparaisons avec un groupe de traitement auquel n'est offert une prime d'encouragement qu'à la première vague.

L'étude a fourni de nouvelles preuves des effets des paiements incitatifs continus dans une étude par panel à plusieurs modes de collecte. Nous avons testé les effets des primes d'encouragement sur l'attrition, la non-réponse partielle ou le biais d'attrition et vérifié si ces effets variaient d'une vague à l'autre. Nous avons vérifié si les primes d'encouragement conditionnelles et inconditionnelles avaient des effets semblables, si les effets des primes différaient selon le mode de collecte et le niveau d'aptitude, et si les changements de traitement incitatif ou de mode de collecte avaient un effet durable sur l'effet des traitements au cours des vagues subséquentes.

Les résultats montrent que les primes d'encouragement inconditionnelles réduisent significativement l'attrition et, quoiqu'elles accroissent aussi la non-réponse partielle, l'effet net sur la quantité d'information recueillie durant l'enquête est positive. Les primes d'encouragement ont des effets proportionnels pour une gamme de caractéristiques des répondants et, par conséquent, n'ont pas d'effet sur le biais d'attrition relatif à ces caractéristiques. La non-réponse partielle augmente davantage avec les primes d'encouragement inconditionnelles que conditionnelles et davantage en cas du mode de collecte par la poste que de celui par téléphone. Le biais d'attrition n'est affecté ni par le traitement incitatif ni par le mode de collecte. D'une vague à l'autre, les primes d'encouragement ont eu un effet quelque peu décroissant sur l'attrition, mais des effets semblables sur la non-réponse partielle. L'absence d'effets sur le biais d'attrition était aussi un élément constant d'une vague à l'autre. Les changements de traitement incitatif pour passer d'une prime conditionnelle à une prime inconditionnelle et le passage du mode de collecte par téléphone au mode de collecte par la poste n'ont pas d'effets sur les résultats observés lors de vagues ultérieures.

Les résultats impliquent que les primes d'encouragement offertes aux répondants constituent un moyen efficace de maintenir la taille de l'échantillon d'un panel et d'être certain de l'efficacité des estimations et de la faisabilité des analyses de sous-groupes. Parmi les élèves moins performants, au moins 50 % d'information supplémentaire a été recueillie durant les trois vagues expérimentales, pour ce qui est du nombre de questions à réponse valide par cas émis au départ. Les primes d'encouragement offertes étaient sûres, en ce sens que l'accroissement des taux de réponse n'a pas augmenté par inadvertance le biais de non-réponse relatif aux caractéristiques observées.

Les changements de traitement incitatif n'ont pas eu d'effet durable; cependant, dans la présente étude, le seul changement appliqué était une amélioration pour le répondant, c'est-à-dire le passage d'une prime conditionnelle à

Tableau 5 Effet net des primes d'encouragement sur la réponse totale et la réponse partielle

Nbre de questions valides par unité émise à la vague 2 : prime d'encouragement/témoin	Nbre moyen prévu de NRP				TR cumulatif prévu (%)				Nbre moyen prévu de NRP			
	v2	v3	v4	v4	v2	v3	v4	v4	v2	v3	v4	v4
Élèves plus performants	P <sub>x</sub>	78,23	57,14	43,47	2,89	2,54	1,75	-	1,75	2,54	1,77	1,395
	P <sub>n</sub>	86,45	70,82	60,66	3,19	2,96	1,77	-	1,097	1,52	1,324	1,395
	P <sub>c</sub>	82,32	66,80	57,91	3,30	3,20	2,01	-	1,042	1,52	1,324	1,395
	T <sub>x</sub>	76,82	53,11	38,31	0,12	0,30	1,61	-	-	-	-	-
	T <sub>n</sub>	80,91	59,20	50,69	0,12	0,25	1,80	-	1,053	1,116	1,317	-
	T <sub>c</sub>	64,93	41,14	26,89	3,49	3,52	2,54	-	-	-	-	-
Élèves moins performants	P <sub>x</sub>	53,95	40,63	4,37	4,51	2,91	1,130	-	1,282	1,498	1,498	1,498
	P <sub>n</sub>	71,35	50,09	40,60	3,83	4,66	2,72	-	1,090	1,186	1,504	1,504
	P <sub>c</sub>	65,21	38,53	24,34	0,50	3,48	2,35	-	-	-	-	-
	T <sub>x</sub>	70,41	36,99	0,49	3,56	2,99	1,080	-	1,294	1,498	1,498	1,498
	T <sub>n</sub>	49,94	36,99	0,49	3,56	2,99	1,080	-	1,294	1,498	1,498	1,498

Nota : TR = taux de réponse, NRP = non-réponse partielle, T = téléphone, P = poste, x = témoin, n = prime d'encouragement inconditionnelle, c = prime d'encouragement conditionnelle. Les calculs sont fondés sur 44 questions ramifiées à la vague 2, 48 à la vague 3 et 46 à la vague 4. Le nombre de questions avec réponse valide est calculé selon l'expression  $TR_x * (44 - NRP_2 + 48 - NRP_3 + 46 - NRP_4)$ .

Tableau 6 Énoncé des questions incluses dans l'analyse du biais de non-réponse

Variable	Énoncé de la question
Résultats des examens de 1 <sup>re</sup> année	« Veuillez nous dire : a) quels sujets du CGES vous avez étudiés en 10 <sup>e</sup> et en 11 <sup>e</sup> années, b) pour quels sujets du CGES vous avez passé un examen et c) vos résultats au CGES (n'inscrivez aucun résultat de reprise obtenu en 1 <sup>re</sup> année). »
Attitudes	« Voici certaines choses que des gens ont dit. Nous aimerions savoir ce que vous en pensez. Veuillez inscrire une croix dans l'une des cases pour chaque énoncé : D'accord, En désaccord. Ne sais pas. »
ATT : employeurs	D'accord : « La plupart des employeurs n'offrent pas aux jeunes le type approprié de formation au travail. »
ATT : formation/remunération	D'accord : « Dans ma recherche d'un emploi, je me préoccupe plus d'en trouver un avec offre de formation que de trouver celui qui paie le plus. »
ATT : plans	D'accord : « Je pense que faire des plans d'avenir est une perte de temps. »
ATT : information	D'accord : « Je sais comment obtenir des renseignements au sujet de futures possibilités de travail, de formation ou d'études. »
ATT : soutien	D'accord : « Je reçois suffisamment de soutien pour planifier mon avenir. »
Activité courante :	« Veuillez inscrire une croix dans une case pour nous indiquer votre activité principale à l'heure actuelle : a) sans emploi/chômeur, b) apprentissage moderne, stage de formation nationale, formation de la jeunesse ou autre formation financée par l'État, c) emploi à temps plein (plus de 30 heures par semaine), d) emploi à temps partiel (s'il s'agit de votre activité principale), e) études à temps plein dans une école ou un collège, f) m'occupe de la maison ou de la famille, g) fais autre chose (veuillez préciser). »
Études t.p.	Études à temps plein.
Emploi t.p.	Emploi à temps plein.
PEEF	Pas d'emploi, d'études ni de formation.
Chômeur selon l'OIT	En chômage et à la recherche d'un travail parmi les personnes économiquement actives (variable dérivée de l'YCS).
Chômeur	En chômage durant un ou plusieurs mois d'avril 1999 à mars 2000 : « Nous aimerions également savoir ce que vous avez fait pendant l'entrevue ou la plupart de chaque mois. »
Études (gén.)	Oui : « À l'heure actuelle-vous pour l'obtention d'un certificat de niveau GCSE, A/S ou A? »
Études (prof.)	Oui : « À l'heure actuelle-vous pour l'obtention de tout certificat NVQ (National Vocational Qualification) ou tout autre certificat d'études professionnelles, y compris les certificats BTEC, City & Guilds ou RSA? »
Ménage	« Qui vit dans le même ménage que vous? a) père, b) beau-père, c) mère, d) belle-mère, e) vos propres enfants, f) frères et sœurs, g) autres personnes (veuillez indiquer le lien avec vous). »
Vit avec les parents	Vit avec une ou plusieurs des personnes suivantes : père, beau-père, mère ou belle-mère.
Vit avec un(e) conjoint(e)	Vit avec son petit ami, sa petite amie, son mari, sa femme ou un(e) conjoint(e).



H5 : Effets des changements de mode de collecte ou de traitement incitatif.

Le changement du traitement incitatif ou du mode de collecte n'a pas eu d'effet durable. Le passage de primes conditionnelles aux primes inconditionnelles n'a eu aucun effet durable sur l'attrition ni sur la non-réponse partielle ( $F > 0,05$  pour les tests de Wald de l'égalité de  $\beta_2$  et  $\beta_3$  pour  $t = 3, 4$ ) et les effets après le changement de traitement étaient semblables à ceux observés pour l'échantillon affecté dès le départ à un traitement par prime d'encouragement incondi- tionnelle (tableaux 2 et 3, colonne 5). Le passage de la collecte par téléphone à la collecte par la poste n'a pas eu d'effet durable sur l'attrition ni sur la non-réponse partielle (valeur  $P$  de l'interaction pour le mode de collecte par téléphone et les primes inconditionnelles  $> 0,05$  à  $t = 4$ ) et les effets après le changement de mode de collecte ne différaient pas de ceux observés pour l'échantillon affecté dès le départ à la collecte par la poste et à des primes d'encouragement incondi- tionnelles (tableaux 2 et 3, colonne 8).

H6 : Effets différentiels selon le niveau d'aptitude.

Les effets des primes d'encouragement étaient les mêmes pour tous les niveaux de performance. Les écarts entre les effets proportionnels des primes inconditionnelles et conditionnelles sur l'attrition et la non-réponse partielle selon le niveau de performance n'étaient pas significatifs (les colonnes 6 et 7 des tableaux 2 et 3 donnent les valeurs  $P$  des interactions au niveau de performance avec chacun des traitements incitatifs), puisque les effets absolus étaient compa- rables. Par exemple, les effets des primes d'encou- ragement inconditionnelles ont réduit l'attrition à la vague 2 de 8 points de pourcentage chez les élèves plus performants et de 10 points de pourcentage chez les élèves moins performants. Toutefois, puisque le niveau de non-réponses dans le groupe témoin était 61 % plus élevé pour le groupe de élèves moins performants, l'effet absolu comparable implique un effet proportionnel plus faible, égal à 76 % seulement de l'effet pour les élèves plus performants. De même, la différence entre les modes de collecte ne variait pas selon la performance (la colonne 9 des tableaux 2 et 3 donne les valeurs  $P$  pour l'interaction entre le niveau de performance, les primes d'encou- ragement inconditionnelles et le mode de collecte par téléphone) et l'absence d'effet sur le biais d'attrition ne différait pas pour les élèves moins performants (données non présentées).

d'encouragement sur le biais d'attrition ne variaient pas d'une vague à l'autre.

H3 : Primes d'encouragement conditionnelles comparati-

vement à inconditionnelles.

Les primes d'encouragement inconditionnelles ont eu un effet de réduction de l'attrition plus prononcé que les primes conditionnelles, mais des effets semblables sur la non-réponse partielle et le biais d'attrition. Dans le cas des élèves les plus performants, les primes conditionnelles utilisées à la vague 2 ont réduit l'attri- tion deux fois moins fort que les primes incondi- tionnelles (tableau 2, colonne 5) et l'écart entre les deux traitements était significatif (la valeur  $P$  d'un test de Wald de l'égalité de  $\beta_2$  et  $\beta_3$  était 0,0142). Simultanément, les primes conditionnelles ont augmenté la non-réponse partielle de 38 % de plus que les primes non conditionnelles (tableau 3, colonne 5), mais l'écart n'était pas significatif. Les primes con- ditionnelles ont réduit quelque peu le biais d'attrition pour une seule caractéristique, à savoir les membres de l'échantillon appartenant au groupe témoin qui étudiaient pour l'obtention d'un certificat d'études professionnelles à la vague 1 étaient 6,8 % plus susceptibles de décrocher de l'enquête que ceux ne faisant pas ce genre d'études. Dans le cas des primes d'encouragement conditionnelles, l'écart était de 6,4 % (la valeur  $P$  de l'interaction des primes conditionnelles avec la caractéristique était  $\leq 0,05$  pour  $t = 2$ ).

H4 : Effets différentiels selon le mode de collecte.

Les primes d'encouragement ont eu plus d'effets sur l'attrition et la non-réponse partielle dans le cas de la collecte par la poste que par téléphone, mais n'ont eu d'effet sur le biais d'attrition ni pour l'un ni pour l'autre mode. Dans le cas du mode de collecte par téléphone, les primes d'encouragement incondi- tionnelles ont eu un effet sur l'attrition égal à moins de la moitié de celui qu'elles ont eu dans le cas du mode de collecte par la poste pour le groupe des élèves plus performants (tableau 2, colonne 8). L'écart était significatif à la vague 3 (la valeur  $P$  de l'interaction entre le mode de collecte par téléphone et les primes d'encouragement inconditionnelles était de 0,0057), mais non à la vague 2. À la vague 3, les primes ont augmenté la non-réponse partielle de 5 % de moins dans le cas de la collecte par téléphone que dans celui de la collecte par la poste (la valeur  $P$  de l'interaction était 0,0442), mais l'écart à la vague 2 n'était pas significatif. Le manque d'effet des primes conditionnelles sur l'attrition ne variait pas selon le mode de collecte.

le biais en fonction des attitudes relatives à la formation (valeur  $P = 0,0269$  pour  $t = 2$ ); inférieur de 4 % à la vague 2 (valeur  $P = 0,00269$  pour  $t = 2, 3, 4$ ). Le biais d'attrition pour le mode de collecte par téléphone ne différait pas de celui pour le mode de collecte par la poste, sauf en cas de non-réponse différentielle selon le sexe : le biais était 7 % plus faible à la vague 2, 2 % plus faible à la vague 3 et 1 % plus élevé à la vague 4 (la valeur  $P$  de l'interaction entre le mode de collecte par téléphone et le sexe était  $\leq 0,002$  pour  $t = 2, 3, 4$ ).

## 6. Évaluation des hypothèses

Les résultats dont nous discutons ici sont résumés au tableau 2 (effets des primes d'encouragement sur l'attrition), tableau 3 (effets sur la non-réponse partielle), tableau 4 (effets sur le biais d'attrition) et tableau 5 (effet net sur la non-réponse totale et la non-réponse partielle).

**H1 : Effets des primes d'encouragement sur le taux d'attrition, le biais d'attrition et la non-réponse partielle.**  
Les primes d'encouragement ont réduit l'attrition et augmenté la non-réponse partielle, mais n'ont pas eu d'incidence sur le biais d'attrition. Les primes d'encouragement incitantes ont réduit l'attrition cumulée dans l'échantillon d'élèves plus performants (ce qui correspond à un écart de 8 points de pourcentage) à la vague 2, de 32 % à la vague 3 et de 30 % à la vague 4 (valeur  $P$  de  $\beta_2 = 0,0000$  pour  $t = 2, 3, 4$ ). Parallèlement, les primes d'encouragement ont augmenté la non-réponse partielle de 10 % à la vague 2 et de 17 % à la vague 3 (valeur  $P$  de  $\beta_2 \leq 0,05$  pour  $t = 2, 3$ ), mais n'ont eu aucun effet à la vague 4 (tableau 3, colonne 4). L'écart d'une vague à l'autre n'était toutefois pas significatif (voir H2). Les primes d'encouragement ont eu un effet proportionnel sur l'attrition pour toutes les caractéristiques des répondants et, par conséquent, n'ont pas réduit le biais d'attrition : la valeur  $P$  de l'interaction des primes d'encouragement incitantes et des caractéristiques des répondants était  $> 0,05$  pour toutes les caractéristiques et les vagues (données non présentées). Faisait exception la proportion d'élèves dans les écoles « modernes » qui étaient sous-représentés dans chacune des trois vagues. (Les écoles

**H2 : Effets des primes d'encouragement d'une vague à l'autre.**

L'effet sur l'attrition a diminué quelque peu d'une vague à l'autre. L'effet relatif de la valeur  $P$  d'un test de Wald de l'égalité de  $\beta_2$  entre les vagues était  $\leq 0,05$ . Alors que l'effet relatif des primes d'encouragement a diminué, l'effet absolu a augmenté d'une vague à l'autre (-17 points de pourcentage à la vague 4, comparativement à -8 et à -14 aux vagues 2 et 3, voir le tableau 5). L'effet des primes d'encouragement sur la non-réponse partielle ne différait pas significativement d'une vague à l'autre (la valeur  $P$  de l'égalité de  $\beta_2$  était  $> 0,05$ ), quoique les nombres prédits de réponses manquantes à une question ont diminué d'une vague à l'autre. De même, les effets des primes

Puisque les primes d'encouragement ont eu un effet positif sur la réponse totale et un effet négatif sur la réponse partielle, le tableau 5 donne l'effet net sur la quantité d'information recueillie dans l'enquête. Les avantages des primes d'encouragement en ce qui concerne la non-réponse totale l'emportent clairement sur le coût en non-réponse partielle. Pour chaque membre de l'échantillon émis à la vague 2, les taux prévus de réponse totale et de réponse partielle pour l'échantillon d'élèves plus performants interviewés par la poste impliquaient qu'à la quatrième vague, 40 % de réponses valides à une question de plus ont été recueillies dans le cas de primes d'encouragement incitantes, 50 % d'information supplémentaire a été recueillie grâce aux primes d'encouragement (effets nets des primes d'encouragement, dans une analyse multivariée ou pour l'analyse des changements, des profils différents de réponse manquante d'une question à l'autre ou d'une vague à l'autre peuvent entraîner la sous-prévision d'un grand nombre de cas par élimination par paire).

L'interaction des primes d'encouragement et de la variable école moderne  $\leq 0,01$  pour  $t = 2, 3, 4$ ).







intervenus par la poste et n'ayant pas reçu de prime d'encouragement constitue la catégorie de référence, à laquelle toutes les autres catégories de traitement sont comparées.

### 5.1 Attrition

Le taux cumulatif prévu d'attrition chez les élèves plus performants affectés au groupe témoin interviewé par la poste est passé de 21,77 % à la vague 2 à 56,53 % à la vague 4 (tableau 2, colonne 1). Pour les *élèves moins performants* (colonne 2), les taux d'attrition dans le groupe témoin interviewé par la poste étaient 61 % plus élevés à la vague 2, mais cet écart s'est amoindri d'une vague à l'autre pour atteindre 29 % à la vague 4. L'écart selon le niveau de performance était néanmoins significatif pour chacune des trois vagues (valeur  $P$  de  $\beta_4 = 0,0000$  pour  $t = 2,3,4$ ). Dans le cas du *mode téléphonique* (colonne 3), le taux d'attrition dans le groupe témoin ne différait pas significativement à la vague 2, mais était supérieur de 9 % à la vague 3 (valeur  $P$  de  $\beta_3 = 0,0034$  pour  $t = 3$ ). Ce résultat est contraire à celui d'autres études, où la non-réponse est généralement plus faible pour le mode d'interview par téléphone, à cause du rôle joué par l'interviewer dans la persuasion des répondants à participer à l'enquête. L'une des raisons possibles du résultat opposé observé ici est qu'ainsi bien pour le groupe interviewé par la poste que pour celui interviewé par ITAO, les efforts supplémentaires en vue d'obtenir les réponses auprès des non-répondants initiaux ont été faits par téléphone, de sorte que seul le groupe interviewé par la poste a été soumis à un traitement multimodes. Deuxièmement, à la vague 2, le fardeau de l'enquête (mesuré par la longueur de l'interview) était plus élevé pour les personnes interviewées par téléphone que celles approchées par la poste à cause des modules supplémentaires, ce qui pourrait avoir donné lieu à une non-réponse plus élevée à la vague 3 chez les premières que chez les secondes. Les taux de réponses cumulatifs prévus, qui ont servi de base pour le calcul des différences en pourcentage entre les groupes de traitement, sont présentés dans les trois premières colonnes du tableau 5.

### 5.2 Non-réponse partielle

Le nombre prévu de réponses manquantes à une question dans le groupe témoin des élèves plus performants interviewés par la poste était de 2,89 pour la vague 2, et tombait à 1,75 pour la vague 4 (tableau 3, colonne 1). Pour les *élèves moins performants* (colonne 2), le nombre prévu pour le groupe témoin était 21 % plus élevé à la vague 2, l'écart se creusant pour atteindre 45 % à la vague 4. Les différences selon le niveau de performance sont significatives pour les trois vagues ( $P \leq 0,0001$  pour  $\beta_4$ ,  $t = 2,3,4$ ). Pour le *mode téléphonique* (colonne 3), le

coefficient pour la caractéristique de répondant,  $\beta_{10}$ , fournit les renseignements sur la direction, l'ordre de grandeur et, en combinaison avec son erreur-type, la signification du biais d'attrition pour le groupe témoin des élèves plus performants interviewés par la poste et n'ayant pas reçu de prime d'encouragement. L'interaction de la caractéristique et les indicateurs de prime d'encouragement renseignent au sujet de la variation du biais d'attrition due aux primes d'encouragement. Nous avons calculé la signification de toutes les interactions présentées ici en suivant les recommandations formulées pour les modèles non linéaires par Norton, Wang et Ai (2004) à l'aide de la commande « predictnl » de la version 9 de Stata.

Les caractéristiques testées étaient le sexe, le type d'école, les résultats d'examen, l'activité courante (études à temps plein, emploi ou pas d'études, d'emploi ni de formation), l'expérience du chômage, la poursuite d'études en vue d'obtenir un certificat d'études professionnelles ou générales, la composition du ménage (vit avec les parents, un conjoint ou une conjointe, ni l'un ni l'autre) et un ensemble de questions attitudinales au sujet de l'emploi et de la formation. Les énoncés de toutes les questions sont présentés au tableau 6. Les caractéristiques choisies sont celles pour lesquelles on pouvait s'attendre à ce que les répondants et les non-répondants diffèrent, d'après des études antérieures de la non-réponse à la YCS ou à d'autres enquêtes et la théorie sur la non-réponse (Groves et Couper, 1998; Lynn, Purdon, Hedges et McAleese 1994).

### 4.4 Résultats présentés

Puisque les coefficients provenant des modèles non linéaires ne peuvent pas être interprétés de manière fondamentale (Long 1997), nous donnons les valeurs prévues fondées sur les estimations du modèle, plutôt que sur les coefficients. Sauf indication contraire, les résultats sont présentés pour le groupe des élèves plus performants. Pour donner une idée de la grandeur des différences de résultats entre les traitements, nous présentons les transformations entre les valeurs prévues, qui comparent chaque traitement avec le groupe de référence, c'est-à-dire le groupe témoin d'élèves plus performants interviewés par la poste.

## 5. Attrition, non-réponse partielle et biais d'attrition dans les groupes témoins

À titre de contexte de l'évaluation des effets des primes d'encouragement, la présente section fournit des renseignements sur la portée de l'attrition, de la non-réponse partielle et du biais d'attrition dans les groupes témoins, en mettant en relief les différences entre les niveaux de performance des élèves et les modes de collecte. Tout au long de la discussion, le groupe des élèves plus performants

Le tableau 1 donne aussi les taux de réponse d'une vague sur l'autre ainsi que cumulatifs observés (RRI de l'AAPOR). Les taux sont présentés selon le niveau de performance des élèves et la combinaison séquentielle de collecte/traitement inchaîné. Les taux de réponse d'une vague sur l'autre pour l'échantillon d'élèves plus performants affectés au groupe témoin interviewé par téléphone à la vague 2 et transférés au groupe témoin interviewé par la vague 4 (colonne 1) étaient, par exemple, 76,82 % et 69,13 % et 72,21 %. Les nombres émis de cas ont diminué, pour passer de 2 075 à 1 101 au cours des trois vagues, parce que les cas non répondants n'ont pas été repris dans l'échantillon lors des vagues subséquentes.

#### 4. Mesures des résultats et méthodes

L'analyse est fondée sur l'échantillon de répondants de la vague 1, puisque l'affectation aux traitements expérimentaux s'appuie sur l'information recueillie au cours de la première vague et que les caractéristiques correspondantes des non-répondants de la vague 1 sont inconnues. Nous nous concentrons par conséquent sur l'attrition, conditionnellement à la réponse à la vague 1. Il s'agit de l'aspect de la non-réponse qui est particulièrement aux enquêtes par panel, bien qu'il faille évidemment reconnaître que les caractéristiques de l'attrition sont conditionnées par celles de la réponse à la vague 1. Le taux de réponse à la vague 1 (RRI de l'AAPOR) était de 54,80 %, après exclusion de cinq cas dont l'admissibilité était connue (Russell et Phelps 2001). La présente section décrit les mesures des résultats et les méthodes utilisées pour évaluer les hypothèses quant aux effets des primes d'encouragement.

##### 4.1 Attrition

Afin de tester l'effet des primes d'encouragement sur l'attrition, nous avons estimé la probabilité d'attrition sous la forme d'une fonction des variables du plan expérimental [mode téléphonique (*tel*), primes d'encouragement inconditionnelles (*unc*), primes d'encouragement conditionnelles (*cond*), élèves moins performants (*ld*) et leurs interactions]. Pour chacune des trois vagues expérimentales ( $t = 2, 3, 4$ ), nous avons estimé un modèle probit distinct de la probabilité d'attrition, en utilisant dans chaque cas l'échantillon de la vague 1 comme base :

$$\Pr(\text{attrition}_t) = F(\beta_{0t} + \beta_{1t}tel_t + \beta_{2t}unc_t + \beta_{3t}cond_t + \beta_{4t}la_t + \beta_{5t}tel_t * unc_t + \beta_{6t}la_t * cond_t + \beta_{7t}la_t * unc_t + \beta_{8t}la_t * cond_t + \beta_{9t}la_t * tel_t * unc_t + e_t) \quad (1)$$

Afin de tester l'effet des primes d'encouragement sur la non-réponse partielle, nous avons estimé des modèles de comptage du nombre de réponses manquantes à une question, en utilisant toutes les questions non filtrées du questionnaire de base utilisé à la vague 2 ( $n = 44$ ), 3 ( $n = 48$ ) et 4 ( $n = 46$ ), où toute réponse « Ne sais pas » a été comptée comme une valeur manquante. Nous avons utilisé la même spécification des prédicteurs que dans le modèle (1) pour estimer des modèles de régression bino-miale négative distincts pour chacune des trois vagues expérimentales, conditionnellement à la réponse à la vague en question. (La surdispersion signifiait que les modèles de Poisson n'étaient pas ajustés aux données : la valeur  $P$  du test du rapport de vraisemblance de même moyenne et de même variance était 0,0000 pour toutes les vagues). Les coefficients et les erreurs-types estimés pour ces modèles ont été utilisés pour calculer la non-réponse partielle prévue sous diverses conditions de traitement et pour tester le degré de signification des différences dues aux primes d'encouragement.

##### 4.2 Non-réponse partielle

L'existence de différences due aux primes d'encouragement sous diverses conditions de traitement et pour tester

Pour tester l'effet des primes d'encouragement sur les biais d'attrition, nous avons estimé la probabilité d'attrition en utilisant le modèle (1), mais en y incluant les caractéristiques des répondants à la vague (1) et leurs interactions avec les variables du plan expérimental comme prédicteurs. Nous avons estimé des modèles probit distincts pour l'attrition pour chacune des vagues expérimentales ( $t = 2, 3, 4$ ) et pour chaque caractéristique, de nouveau en utilisant l'échantillon de répondants de la vague 1 (*vt*) comme base :

$\Pr(\text{attrition}_t) = F(\beta_{0t} + \beta_{1t}tel_t + \beta_{2t}unc_t + \beta_{3t}cond_t + \beta_{4t}la_t + \beta_{5t}tel_t * unc_t + \beta_{6t}la_t * tel_t + \beta_{7t}la_t * unc_t + \beta_{8t}la_t * cond_t + \beta_{9t}la_t * tel_t * unc_t + \beta_{10t}v1char_t + \beta_{11t}v1char_t * tel_t + \dots) \quad (2)$

où  $\beta_{11}$  à  $\beta_{19}$  sont les coefficients des termes d'interaction des caractéristiques avec les variables du plan. Le



était de 1/2 si le ratio d'élèves moins performants aux élèves plus performants dans l'échantillon était  $\geq 2$ , de 1/3 si  $1/2 \leq \text{ratio} < 2$  et de 1/4 pour toutes les autres écoles.) La procédure a été répétée pour l'affectation des écoles dans les groupes de traitement par la poste, où celles sélectionnées pour un traitement incitatif ont été réparties aléatoirement en un groupe de traitement conditionnel et un groupe de traitement inconditionnel. (La proportion affectée au traitement inconditionnel était de 2/3 si le ratio était  $\geq 2$ ; de 1/3 pour  $0 \leq \text{ratio} < 1/2$  et de 1/6 pour toutes les autres écoles.) Toutes les estimations du degré de signification présentées dans le texte tiennent compte du plan d'échantillonnage par grappe de l'expérience sur les primes d'encouragement.

Le tableau 1 montre les tailles des échantillons émis sur le terrain à chaque vague pour les diverses combinaisons de traitement incitatif et de mode de collecte, après exclusion des cas dont l'inadmissibilité était connue, à cause d'un déménagement à l'étranger ou du décès ( $n = 13$  à la vague 2,  $n = 3$  à la vague 3). Les cas inadmissibles à la vague 4 ne sont pas identifiés dans les données, mais leur nombre est vraisemblablement faible. Sont également exclus de l'analyse les répondants de la vague 1 pour lesquels aucun numéro de téléphone n'était connu au moment de l'affectation aux modes de collecte pour la vague 2, comme il est décrit plus haut, et 117 élèves plus performants affectés au mode de collecte par téléphone à la vague 2 qui ont répondu par la poste et ont été subse-

quemment affectés au mode de collecte par la poste.

Tableau 1 Taux conditionnels et cumulatifs de réponse

Vague	Taux de réponse %	Élèves plus performants					Élèves moins performants				
		$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$	$T^xT^xP^x$
2	Conditionnel (n émis) (2,075)	76,82	80,91	78,23	86,45	82,32	65,21	70,41	64,93	75,00	71,35
3	Conditionnel (n émis) (2,075)	69,13	73,17	73,07	81,91	(967)	(1,282)	(811)	(807)	(608)	(569)
	Conditionnel (n émis) (1,594)	53,11	59,20	57,16	70,82	(794)	(836)	(571)	(524)	(456)	(406)
4	Cumulatif (n émis) (1,101)	72,21	85,61	76,11	85,65	(645)	(494)	(404)	(332)	(328)	(284)
	Cumulatif (n émis) (431)	50,69	43,48	60,66	58,03	24,34	36,99	26,89	40,63	40,60	40,60

Nota : Taux de réponse selon la définition normalisée RRI de l'AAPOR. Les groupes de traitement sont désignés par  $T^x$  = téléphone,  $P^x$  = poste,  $x$  = témoin,  $n$  = prime d'encouragement inconditionnelle,  $c$  = prime d'encouragement conditionnelle. Par exemple,  $T^xT^xP^x$  indique que l'échantillon a été affecté au groupe témoin interviewé par téléphone aux vagues 2 et 3, et au groupe témoin interviewé par la poste à la vague 4. Les taux de réponse conditionnels sont conditionnés sur la réponse à la vague précédente. La base du calcul est le nombre de cas émis dans l'échantillon, qui exclut les non-répondants à la vague précédente et les cas inadmissibles. Les taux de réponse cumulatifs correspondent au pourcentage de répondants de la vague 1 qui restent dans l'échantillon de répondants. La base pour le calcul est le nombre de cas émis dans l'échantillon de la vague 2, à l'exclusion de trois élèves plus performants inadmissibles à la vague 3 (1 cas  $P^xP^xP^x$  et 2 cas  $P^cP^xP^x$ ).

pas été repris lors des vagues subséquentes et, par conséquent, l'attrition était monotone.

La vague 1 était une enquête par la poste avec suivi par téléphone des non-répondants après quatre envois par la poste (envoi du questionnaire initial et de trois rappels).

D'après les résultats d'examen qu'ils ont déclarés, les répondants de la vague 1 ont été classés dans la catégorie des « élèves plus performants » s'ils avaient été reçus à cinq examens pour l'obtention du certificat général d'études secondaires (GCSE) avec une note allant de A\* à C, ou dans la catégorie des « élèves moins performants », autrement, l'environ le tiers des répondants de la vague 1 ont été classés ainsi dans la catégorie des moins performants. À l'occasion de la vague 2, environ le tiers des membres de l'échantillon utilisé sur le terrain ont été sélectionnés au hasard pour répondre à des questions supplémentaires sur des sujets particuliers et assignés à des interviews téléphoniques assistées par ordinateur (TTAO). En plus du questionnaire de base, les jeunes interviewés par téléphone ont répondu à un module sur les décisions relatives à la poursuite d'études supérieures (pour les élèves plus performants) ou sur les aspirations en matière d'études et d'emploi (pour les élèves moins performants). Les autres membres de l'échantillon ont reçu le questionnaire de base par la poste. À la vague 3, tous les élèves moins performants ont reçu le questionnaire de base envoyé par la poste, mais le module téléphonique a continué d'être administré à un tiers des élèves plus performants. À la quatrième vague, tous les répondants ont été affectés à l'enquête de base effectuée par la poste. La figure 1 illustre l'affectation aux divers modes de collecte et aux diverses primes d'encouragement.

Le questionnaire de base est demeuré en grande partie le même pour les trois vagues expérimentales. Les jeunes interviewés par téléphone ont d'abord répondu aux questions de base, puis aux modules supplémentaires. Le questionnaire de base était le même que le questionnaire envoyé par la poste, quoique certaines questions ont été adaptées en vue de leur administration par téléphone. L'interview téléphonique durait, en moyenne, 20 minutes. (Les questionnaires et les rapports techniques peuvent être consultés par

Questionnaire		De base + études supérieures		De base		De base + études et emplois		De base	
Vague 4	--	--	--	$P^x$	$P^u$	--	--	$P^x$	$P^u$
Vague 3	$T^x$	$T^u$	$T^u$	$P^x$	$P^u$	--	--	$P^x$	$P^u$
Vague 2	$T^x$	$T^u$	$T^u$	$P^x$	$P^u$	$T^x$	$T^u$	$P^x$	$P^u$
Elèves plus performants					Elèves moins performants				
les répondants pour l'affectation à la vague 2.									
Questionnaire de base envoyé par la poste. Résultats d'examens déclarés utilisés pour classer									
Vague 1									

Nota : T = téléphone, P = poste, x = témoin, u = prime inconditionnelle, c = prime conditionnelle. Les flèches indiquent les changements de traitement incitatif ou de mode de collecte d'une vague à l'autre.

Figure 1 Plan expérimental

En réponse aux préoccupations croissantes suscitées par la baisse des taux de réponse, une expérience a été introduite dans la deuxième vague de la cohorte 10 afin d'étudier l'effet de paiements d'encouragement sur les taux de réponse et sur le biais de non-réponse. Un bon de 5 livres sterling (environ 10 \$US ou 7 euros) a été envoyé à une partie des répondants de la vague 1 de l'enquête réalisée par la poste ainsi que de celle réalisée par téléphone, tandis que les groupes témoins n'ont pas reçu de prime d'encouragement. En outre, dans le cas de l'enquête par la poste, les primes d'encouragement étaient inconditionnelles (la prime a été envoyée avec l'envoi par la poste initial) ou conditionnelles (le bon a été promis dans l'envoi par la poste original, mais envoyé uniquement après réception du questionnaire dûment rempli). À l'occasion des vagues 3 et 4, toutes les primes d'encouragement ont été payées inconditionnellement.

### 3.3 Affectation des répondants aux modes de collecte et aux traitements incitatifs

À l'occasion de la vague 2, les répondants de la vague 1 ont été affectés aléatoirement au mode de collecte par téléphone ou par la poste. Par contre, l'affectation des traitements incitatifs a eu lieu au niveau de l'école (affectation randomisée par grappe selon le mode). Chaque école représentée dans l'échantillon a été affectée à un traitement par téléphone (témoin ou prime d'encouragement) et affectée indépendamment à un traitement, éventuellement différent, par la poste (témoin, prime d'encouragement inconditionnelle ou prime d'encouragement conditionnelle), de sorte que tous les membres de l'échantillon appartenant à une même école et approchés selon le même mode de collecte ont reçu le même traitement incitatif.



H6 : Effets des primes d'encouragement selon le niveau d'aptitude.

Les membres de l'échantillon dont le niveau d'études est faible sont habituellement plus susceptibles de décrocher des enquêtes. Si les primes d'encouragement réduisent le biais d'attrition, elles devraient par conséquent réduire de manière disproportionnée l'attrition chez les personnes moins performantes. Les répondants dont le niveau d'aptitude est faible pourraient aussi être plus susceptibles de fournir des réponses incomplètes, s'ils trouvent plus difficile que les autres la tâche de répondre au questionnaire envoyé par la poste. Donc, les primes d'encouragement peuvent accroître les niveaux moyens de non-réponse partielle.

Hypothèse nulle H6 : Les primes d'encouragement ont les mêmes effets quel que soit le niveau d'aptitude.

### 3. Plan de l'étude

La Youth Cohort Study of England and Wales (YCS) a pour but d'étudier les transitions de la scolarité obligatoire à la poursuite d'études de niveau plus élevé ou au marché du travail et, en général, comprend l'échantillonnage tous les deux ans de cohortes de jeunes de 16 et 17 ans qui sont étudiées à plusieurs occasions à intervalle d'un an. L'expérience sur les primes d'encouragement a été intégrée dans les vagues 2, 3 et 4 de la cohorte 10. L'enquête est gérée et financée par le Department for Children, Schools and Families, qui a conçu l'expérience sur les primes d'encouragement en collaboration avec le National Centre for Social Research, qui a réalisé en sous-traitance les vagues 2 et 3 de la cohorte 10 de la YCS.

#### 3.1 L'enquête

La population étudiée dans la cohorte 10 de l'YCS comprenait les élèves d'Angleterre et du pays de Galles ayant atteint l'âge minimal pour quitter l'école (16 ans) durant l'année scolaire 1998-1999 (Russell et Phelps 2001), c'est-à-dire une cohorte d'une année d'élèves nés entre le 1<sup>er</sup> septembre 1982 et le 31 août 1983. Un échantillon aléatoire au 1/10 a été tiré des registres des écoles (à l'exclusion des écoles spéciales et des écoles comptant moins de 20 élèves de cet âge) en 1999, en demandant aux écoles de fournir les noms et adresses des élèves nés le 5, le 15 et le 25 de chaque mois. Puis, un échantillon aléatoire systématique de 25 000 élèves a été tiré du fichier résultant contenant 31 424 noms et adresses. La première vague de l'enquête a eu lieu un an plus tard, au printemps 2000, la deuxième à la fin de 2000, la troisième au printemps 2002 et la quatrième au printemps 2003. Les non-répondants n'ont

H3 : Effets des primes d'encouragement inconditionnelles et conditionnelles dans le contexte d'un panel.

Selon des études antérieures portant sur des enquêtes transversales, les primes d'encouragement inconditionnelles ont des effets plus importants sur la non-réponse totale, peut-être parce que le paiement préalable signalé que l'organisme d'enquête pense qu'il peut compter sur la participation des membres de l'échantillon, ce qui renforce la norme de réciprocité. Il reste à déterminer si les diverses conditions incitatives ont des effets différents sur la non-réponse partielle.

Hypothèse nulle H3 : Les primes d'encouragement inconditionnelles et conditionnelles ont les mêmes effets dans un contexte de panel.

H4 : Effets des primes d'encouragement dans les modes de collecte par la poste et par téléphone.

Les comparaisons d'études antérieures donnent à penser que les primes d'encouragement ont un effet plus important lorsque la collecte de données se fait par la poste, peut-être parce que dans le mode de collecte par téléphone, l'interv intervieweur agit déjà comme un motivateur externe accroissant la réponse totale ainsi que partielle (Singer et coll. 1999) et que l'importance des améliorations supplémentaires est plus faible. Il pourrait ne pas en être de même dans le contexte d'un panel, où l'effet du mode de collecte sur la réponse peut être modifié par l'expérience du répondant lors des vagues antérieures.

Hypothèse nulle H4 : Les primes d'encouragement ont les mêmes effets en mode de collecte par la poste et par téléphone.

H5 : Effets des changements de mode de collecte ou de traitement incitatif d'une vague de l'enquête à l'autre.

Comparativement aux membres de l'échantillon affectés au même mode de collecte et au même traitement à toutes les vagues, ceux qui sont affectés à des traitements différents ou à des modes de collecte différents ont peut-être vécu des expériences différentes durant les vagues précédentes de l'enquête et ont peut-être des attentes différentes au sujet des vagues futures. Par conséquent, l'effet des primes d'encouragement pourrait non seulement être conditionnel au mode de collecte durant la vague courante, mais être aussi influencé par le traitement incitatif et le mode de collecte des vagues antérieures.

Hypothèse nulle H5 : Les changements de mode de collecte ou de traitement incitatif entre les vagues n'ont pas d'effets persistants.



des résultats sont utilisées pour tester les hypothèses suivantes :

H1 : *Effets des primes d'encouragement sur l'attrition, la non-réponse partielle et le biais d'attrition.*

Dans le cadre d'études antérieures, les primes d'encouragement faisaient généralement croire les taux de réponse, que ce soit parce que les normes d'échanges sociaux obligent les répondants à rendre la pareille (norme de réciprocité, Gouldner 1960) ou parce que la prime d'encouragement se substitue au manque de motivation à participer pour d'autres raisons, comme le devoir civique ou l'intérêt pour le sujet [théorie du levier et de la saillance (*leverage-salience theory*), Groves, Singer et Comins 2000]. En plus de motiver les répondants à fournir des réponses de meilleure qualité, les primes d'encouragement peuvent réduire la non-réponse partielle. Parallèlement, elles peuvent modifier la composition de l'échantillon en y intégrant un plus grand nombre de répondants ne manifestant pas de diligence à répondre aux questions de l'enquête et, par conséquent, accroître la non-réponse partielle. Enfin, les primes d'encouragement peuvent avoir des effets différents sur l'attrition selon les membres de l'échantillon. En principe, ceux qui ont une forte propension à participer à l'enquête sans la prime devraient être moins affectés par l'offre de primes d'encouragement que ceux qui sont plus susceptibles de décrocher de l'enquête. Par conséquent, les primes d'encouragement pourraient réduire le biais d'attrition. *Hypothèse nulle H1 : Les primes d'encouragement n'ont pas d'effet sur l'attrition, la non-réponse partielle ni le biais d'attrition.*

H2 : *Effets des primes d'encouragement d'une vague à l'autre.*

L'effet d'accroissement des taux de réponse totale et de réponse partielle des primes d'encouragement pourrait s'affaiblir d'une vague à l'autre, si l'attrition se traduit par l'abandon des membres les moins motivés à participer et que les membres restants sont suffisamment motivés pour d'autres raisons et donc moins sensibles aux primes d'encouragement (Laurie et Lynn, sous presse). Cependant, la portée de la réduction du biais de non-réponse pourrait croître au fil des vagues de l'enquête si les primes d'encouragement retiennent de manière disproportionnée les membres de l'échantillon les plus susceptibles d'abandonner. *Hypothèse nulle H2 : Les effets des primes d'encouragement ne varient pas d'une vague à l'autre.*

## 2. Hypothèses testées

parce qu'elles accroissent marginalement la motivation de répondants qui, autrement, auraient cessé de participer à l'étude et qui ne sont pas suffisamment aptes ou motivés pour répondre diligemment, les auteurs des études existantes concluent soit que les primes suscitent un plus grand effort de la part du répondant et produisent une non-réponse partielle plus faible (James et Bolesin 1990; Mack et coll. 1998; Singer et coll. 2000), soit qu'il n'existe aucune association (Berk, Mathiowetz, Ward et White 1987; Davern, Rockwood, Sherrod et Campbell 2003; Goyder 1994; Shettle et Mooney 1999; Singer et coll. 1999; Teisl, Roe et Vayda 2005; Tzamourani et Lynn 1999; Williamson, Schuman, Pennell et Lepkowski 1995). La non-réponse partielle est un facteur qui peut être crucial, parce que les analystes n'utilisent habituellement que les cas pour lesquels les données sont complètes, ce qui entraîne une perte d'efficacité due à la réduction de la taille des échantillons et, comme dans le cas de la non-réponse totale, peut biaiser les estimations et invalider l'inférence si les personnes qui ne répondent pas à certaines questions ne représentent pas un sous-ensemble aléatoire de l'échantillon (Mason, Lessner et Traugott 2002). Les problèmes que pose la non-réponse partielle deviennent plus épineux dans le cas de l'analyse multivariée si les profils de données manquantes varient d'une question à l'autre, ainsi que dans le cas de l'analyse des changements, qui, en plus, dépend de renseignements complets à divers points dans le temps. Puisque les primes d'encouragement peuvent avoir une incidence à la fois sur la non-réponse totale et la non-réponse partielle, leur effet éventuel sur des mesures répétées dérivées d'une étude par panel n'est pas clair. Nous avons examiné l'effet des primes d'encouragement sur les taux de non-réponse partielle et calculé leur effet net sur l'attrition et la réponse partielle.

Les résultats mesurés dans le cadre de la présente analyse sont le taux d'attrition, le taux de non-réponse partielle et le biais d'attrition. Ici, l'attrition est un état sans retour, car l'enquête ne comportait pas la reprise des non-répondants dans l'échantillon à l'occasion de vagues ultérieures. La non-réponse partielle est mesurée par le nombre de questions non filtrées pour lesquelles la réponse manquait, à cause d'un refus ou d'une réponse « Ne sais pas ». (Les questions non filtrées sont celles qui s'appliquent à tous les membres de l'échantillon; les questions pour lesquelles l'admissibilité est déterminée par la réponse à une question antérieure sont exclues de notre mesure de la non-réponse partielle.) Le biais d'attrition est mesuré en fonction des caractéristiques sociodémographiques et des mesures des variables étudiées de la première vague. Ces trois mesures

Marians (2006) fait exception. Les auteurs comparent les effets de primes monétaires et de cadeaux dans une enquête à mode de collecte mixte par la poste et sur place. Néanmoins, leur étude ne comporte pas de scénario sans prime et ne permet donc pas d'évaluer l'ordre de grandeur des effets des primes d'encouragement entre les modes. Nous avons comparé les effets de primes d'encouragement dans une enquête à mode de collecte mixte par la poste et par interview téléphonique assistée par ordinateur (ITAO) et également examiné les effets de primes conditionnelles et inconditionnelles.

La recherche sur les effets des primes d'encouragement s'est concentrée sur les taux de réponse et l'on en sait fort peu au sujet des effets sur le biais, lequel est la raison pour laquelle les faibles de taux de réponse sont préoccupants. Les études sur les primes d'encouragement sont limitées principalement à l'examen des effets sur le biais de composition de l'échantillon et certaines ont révélé que ces primes accroissent de manière disproportionnée la participation des répondants habituellement sous-représentés, comme les personnes peu instruites (Singer, Van Hoewyk et Maher 2000), les pauvres (James 1997), les Noirs ou les pauvres (Mack, Huggins, Kaathley et Sundukchi 1998), les membres de groupes ethniques minoritaires noirs ou indiens, vivant dans de grands ménages ou dans des ménages avec enfants à charge, âgés de 0 à 20 ans, ou les personnes seules (Stratford, Simmonds et Nicolaas 2003). Cependant, les biais de composition de l'échantillon ne sont pas nécessairement corrélés aux biais qui entraînent d'importantes estimations par sondage, notamment parce qu'en général, quelques caractéristiques compositionnelles sont étudiées. En dernière analyse, les études du biais de non-réponse sont restreintes par le manque d'information au sujet des non-répondants, limite que permettent de surmonter en partie les études par panel, où l'information sur les non-répondants peut être extraite des vagues qui ont précédé le décrochage de l'enquête. Nous estimons l'importance du biais dû à l'attrition en fonction de la composition de l'échantillon et des variables étudiées. Puis, nous évaluons l'efficacité avec laquelle les primes d'encouragement réduisent le biais, en exploitant l'information disponible dans le panel sur les personnes qui cessent de participer à l'enquête.

En outre, les renseignements sont peu abondants sur les effets des primes d'encouragement d'une vague à l'autre d'une enquête par panel, que le même traitement soit répété ou qu'il change d'une vague à l'autre. Dans une étude de l'utilisation de primes d'encouragement dans les enquêtes longitudinales, Laune et L'ynn (sous presse) concluent que, compte tenu des répercussions de la modification de

conditions d'encouragement sur les coûts, il existe étonnamment peu de données sur les effets de long terme de tels changements en vue d'orienter les praticiens des sondages. Citant un mémoire interne du US Census Bureau, Ward, Bogges, Selva et McMahon (2001) écrivent (voir la page 2) qu'un examen des études longitudinales bien connues (Downs 1999) a révélé que toutes les études non réalisées par le Census Bureau comportaient l'offre d'une prime monétaire à chaque vague, mais qu'aucun test scientifique n'avait été réalisé en vue de déterminer l'efficacité de ces primes. Si l'attribution correspond au décrochage des participants les moins coopérants, l'échantillon pourrait compter de plus en plus de répondants déterminés à participer, donc moins sensibles aux primes d'encouragement, parce qu'ils sont suffisamment motivés pour participer même sans les primes (Laurie et Lynn, sous presse). Dans ces conditions, l'effet marginal des primes sur le taux de réponse pourrait décroître au cours de la vie du panel. De même, l'effet sur le biais d'attrition pourrait croître, si les primes ont un effet disproportionné sur les membres de l'échantillon qui, en leur absence, seraient plus susceptibles de décrocher. Bien que quelques études sur les primes d'encouragement aient été réalisées dans le contexte de panels, elles ne portaient pour la plupart que sur une seule vague ou examinaient des changements de traitement incitatif à la suite. À la suite, Abreu et Winters (2001) et Ward et coll. (2001), par exemple, ont étudié les effets des primes d'encouragement sur les taux de conversion des non-répondants aux vagues antérieures; Rodgers (2002) et Laurie (2007) ont examiné les effets des changements de valeur des primes dans un panel. Les seules études s'intéressant aux effets des primes d'encouragement sur plus de deux vagues semblent être celles de James (1997), de Mack, Huguenin, Kearley et Sundkuchi (1998) et de Laurie et Lynn (sous presse), qui ont signalé que l'effet positif d'un paiement incitatif au début d'un panel persiste pendant plusieurs vagues, même si le paiement n'est pas répété. Ces études ne portent toutefois que sur l'effet d'un paiement incitatif lors d'une seule vague et n'examinent pas les effets cumulatifs de primes d'encouragement offertes au cours de plusieurs vagues successives. Nous examinons ici les effets cumulatifs de paiements incitatifs continus au cours de trois vagues s'étendant sur une période de trois ans, ainsi que les effets du passage d'un mode de collecte par téléphone à un mode de collecte par la poste et du passage d'un traitement incitatif conditionnel à un traitement incitatif inconditionnel.

Enfin, on trouve dans la littérature des données contradictoires quant aux effets des primes d'encouragement sur la qualité des données. Bien que soient fréquemment exprimées des craintes que les primes d'encouragement aboutissent à la collecte de données de moins bonne qualité



# Offre de primes d'encouragement aux répondants dans une enquête par panel multimodes : effets cumulatifs sur la non-réponse et le biais

Annette Jäckle et Peter Lynn<sup>1</sup>

## Résumé

L'offre de primes en vue d'encourager la participation aux enquêtes est un moyen utilisé de plus en plus fréquemment pour combattre la chute des taux de réponse et les risques résultant de biais dû à la non-réponse. Cette dernière pose tout spécialement un problème dans les enquêtes par panel, car même un faible taux de non-réponse d'une vague à l'autre peut entraîner des pertes cumulées importantes et, si la non-réponse est différentielle, cette situation peut aboutir à l'accroissement du biais d'une vague à l'autre. Les effets des primes d'encouragement ont été bien étudiés dans le contexte d'enquêtes transversales, mais l'on en sait fort peu quant à leurs effets cumulés sur l'ensemble des vagues d'un panel. Nous fournissons de nouvelles preuves des effets des paiements incitatifs continus sur l'attrition, ou érosion de l'échantillon, sur le biais et sur la non-réponse partielle en nous appuyant sur les données d'une expérience d'offre de primes d'encouragement à grande échelle, multivagues, à mode d'interview mixte, intégrée dans une enquête par panel réalisée auprès des jeunes par le gouvernement britannique. Dans cette étude, les primes d'encouragement ont donné lieu à une réduction significative de l'attrition, qui compensait de loin les effets négatifs sur les taux de réponse partiels en ce qui concerne la quantité d'information recueillie dans le cadre de l'enquête par cas émis. Les primes d'encouragement ont eu des effets proportionnels sur les taux de rétention pour une gamme de caractéristiques des répondants et, par conséquent, n'ont pas réduit le biais d'attrition relatif à ces caractéristiques. Les effets des primes d'encouragement sur le taux de rétention étaient plus importants pour les primes inconditionnelles que pour les primes conditionnelles, et plus importants également pour le mode de collecte par la poste que pour celui par téléphone. D'une vague à l'autre, les effets sur l'attrition ont diminué quelque peu, mais les effets sur la non-réponse partielle et l'absence d'effet sur le biais sont demeurés constants. Les effets aux primes et des modes de collecte des données appliqués lors de vagues plus tardives semblent avoir été indépendants des traitements relatifs des primes d'encouragement offertes lors de vagues plus tardives.

Mots clés : Attrition; non-réponse partielle; enquête par la poste; ITAO; prime d'encouragement inconditionnelle; prime d'encouragement conditionnelle.

## 1. Introduction

Partout dans le monde, les taux de réponse aux enquêtes sont à la baisse, ce qui accroît les risques de biais de non-réponse si les caractéristiques des non-répondants diffèrent systématiquement de celles des répondants. Dans le cas des enquêtes par panel, cette situation pose un problème particulier, car même de faibles taux de non-réponse lors de chaque vague peuvent donner lieu à des pertes cumulées importantes. Si la non-réponse est différentielle, le biais pourrait augmenter avec la durée du panel. Afin de relever les taux de participation, les organismes d'enquête offrent de plus en plus fréquemment des primes d'encouragement aux répondants. Le présent article fournit de nouvelles preuves des effets cumulatifs des primes d'encouragement sur l'attrition, le biais d'attrition et le biais de non-réponse partielle, en s'appuyant sur des données provenant d'une expérience d'offre de primes d'encouragement à grande échelle, multivagues, à mode de collecte mixte, intégrée dans une enquête par panel menée auprès des jeunes par le gouvernement britannique.

Les effets des primes d'encouragement ont été étudiés dans de nombreuses circonstances. Ainsi, les incitations monétaires accroissent plus la réponse que les cadeaux et les loteries (Church 1993; Singer, Hoewyk, Gebler, Raghunathan et McGonagle 1999), les primes d'encouragement inconditionnelles (c'est-à-dire, celles qui sont données au moment de la demande de participation à l'enquête) accroissent plus la réponse que les primes conditionnelles (celles qui sont promises en échange de la participation) (Church 1993; Goyder 1994; Hopkins et Guillekson 1992; Singer et coll. 1999), les taux de réponse augmentent avec la valeur de la prime (Armstrong 1975; Church 1993; Fox, Crask et Kim 1988; Hopkins et Guillekson 1992; Rodgers 2002; Yu et Cooper 1983), les primes ont des effets plus importants dans les études où les taux de réponse sont faibles, ainsi que dans les enquêtes avec envoi du questionnaire par la poste que dans celles où le questionnaire est administré par un intervieweur (Singer et coll. 1999). Toutefois, la plupart des données révélant des différences d'effet des primes d'encouragement selon le mode de collecte des données émanent de comparaisons entre études distinctes et ne tiennent pas compte de l'effet des différences entre d'autres mesures qui ont une incidence sur la réponse. Par conséquent, les écarts constatés entre les effets des primes d'encouragement ne sont pas nécessairement de





- Sæmstad, C.-E., Swensson, B. et Wretling, J. (1992). *Model Assisted Survey Sampling*. Springer.
- Stahel, W. (1981). *Robuste Schätzungen. Infinitesimale optimalität und Schätzungen von Kovarianzmatrizen*. Thèse de Doctorat, Swiss Federal Institute of Technology.
- Venables, W.N., et Ripley, B.D. (2002). *Modern Applied Statistics with S* (quatrième éd.). New York : Springer. ISBN 0-387-95457-0.
- Wilks, S.S., et Gnanadesikan, R. (1964). Graphical methods for internal comparisons in multiresponse experiments. *Annals of Mathematical Statistics*, 35, 623-631.





Le tableau 7 montre le nombre d'observations aberrantes détectées. GIMCD détecte 31 des 32 observations aberrantes supplémentaires, tandis que BEM, BEMw, TRC et TRCw en détectent un grand nombre, mais pas toutes. ER et ERw détectent un moins grand nombre d'observations aberrantes supplémentaires. Les observations aberrantes de base pondérées sont toutes détectées par ER et ERw, BEMw et TRCw. Les versions non pondérées de BEM et de TRC détectent un moins grand nombre d'observations aberrantes de base pondérées et GIMCD n'en détecte que quatre. Pondérées ou non, BEM et TRC sont les algorithmes qui détectent le mieux les observations aberrantes de base non pondérées.

La dernière ligne du tableau 7 montre que l'imputation non robuste de GIMCD permet de désigner plus d'observations aberrantes avec valeurs manquantes que les autres méthodes dont les imputations sont déjà robustifiées avant la phase de détection.

Tableau 7 MU281 avec valeurs manquantes et contamination supplémentaire modérée

Méthode	N <sup>br</sup> <sub>com</sub> GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
OA	31	19	6	27	27	28	27
Additionnelles	32	4	20	20	11	20	12
De base	20	4	20	20	11	20	12
De base non pondérées	25	13	15	12	23	19	23
Complètes	24	34	43	38	40	40	40
N <sup>br</sup> OA : nombre d'observations aberrantes, n <sup>br</sup> complètes : nombre d'observations complètes parmi les observations aberrantes détectées.							

Afin de déterminer la défaillance des méthodes quand un nombre élevé d'observations aberrantes supplémentaires contaminent les données, nous établissons que des observations sont aberrantes si (LABEL mod 2 = 0 et pop75 > 20) ou (LABEL mod 3 = 1 et pop75 < 20). À part les observations qui sont déjà des observations aberrantes de base, cela fait 98 observations supplémentaires. Donc, regroupées avec les 25 observations de base non pondérées, elles donnent 43,8 % d'observations aberrantes. Le seuil pour les méthodes est par conséquent fixé à la 123<sup>e</sup> distance de Mahalanobis la plus grande. Le tableau 8 montre qu'à cause de ce seuil élevé, toutes les observations aberrantes de base pondérées sont détectées par toutes les méthodes. Les méthodes GIMCD, ER et ERw ne peuvent faire face au taux élevé d'observations aberrantes. La méthode BEM détecte la plupart des observations aberrantes, suivie de près par BEMw et TRC. La méthode TRCw détecte un peu moins d'observations aberrantes de base non pondérées et d'observations aberrantes supplémentaires.

## 6. Conclusion

Tableau 8 MU281 avec valeurs manquantes et forte contamination supplémentaire									
Méthode	N <sup>br</sup> <sub>com</sub>	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw	
OA	98	20	19	37	85	85	80	80	
Additionnelles	20	20	20	20	20	20	20	20	
De base	20	20	20	20	20	20	20	20	
De base non pondérées	25	21	19	17	23	18	18	13	
Pondérées									

L'algorithme EM pour données normales multivariées peut être adapté à un contexte d'échantillonnage. L'algorithme BACON protège l'algorithme EEM non robuste contre les observations aberrantes quand ce dernier est appliqué dans une itération de l'algorithme BACON. L'algorithme ER utilise une robustification à l'intérieur de l'algorithme EM. Les applications révèlent que cette robustification pourrait ne pas suffire. Toutefois, une explication possible pourrait être le point de départ non robuste de l'étape M dans l'algorithme ER.

La méthode de GIMCD, qui est un algorithme EM non robuste suivi d'une imputation et d'une détection avec l'estimateur de covariance MCD, donne de remarquables bons résultats dans des conditions modérées de valeurs manquantes et de contamination. Sa variabilité quand le taux de valeurs manquantes est élevé est un inconvénient. Des solutions plus stables tenant également compte du plan d'échantillonnage devraient être étudiées. L'algorithme BACON-BEM manifeste de très bonnes capacités de détection, en particulier quand le taux de valeurs manquantes et le taux de contamination sont élevés.

Malgré sa simplicité, l'algorithme TRC est une bonne méthode dans de nombreuses circonstances. Son principal inconvénient semble être l'imputation ponctuelle avec une seule covariable, qui peut poser un problème quand le taux de valeurs manquantes est élevé.

Afin de trouver un bon modèle pour la population, il est important d'utiliser les poids d'échantillonnage. Néanmoins, il est conseillé d'utiliser aussi une version non pondérée et d'examiner les différences. Il se pourrait que les observations aberrantes de base pondérées soient masquées par les poids d'échantillonnage élevés, parce qu'elles peuvent alors dominer dans l'estimation du modèle.

## Remerciements

Le projet de recherche EUREDIT faisait partie du Programme de la technologie de la société d'information (TSI) du cinquième programme cadre de l'Union européenne. La participation de la Suisse à EUREDIT a été

est habituellement plus grand pour les algorithmes pondérés. Dans l'ensemble, l'introduction de valeurs manquantes BEM et TRC, tandis qu'elle affecte moyennement celles de la méthode GIMCD.

Tableau 6 Ensemble de données MU281 avec valeurs manquantes, nombre d'observations aberrantes de base détectées

Méthode	GIMD	GIMCD	ER	ERw	BEM	BEMw	TRC	TRCw
Pondérées	14	4	10	20	9	19	12	17
Non pondérées	12	4	23	11	22	15	22	17
Complètes	16	8	17	20	16	18	19	18

GIMD : Imputation non robuste sous un modèle gaussien suivie d'une détection par distance classique de Mahalanobis.

### 5.2.2 Observations aberrantes supplémentaires

En plus des observations aberrantes dans les données originales, nous en introduisons maintenant de nouvelles. Les observations qui devraient devenir des observations aberrantes supplémentaires sont déterminées par le modulo de l'étiquette originale. Si (LABEL mod 8 = 1 et  $\text{pop75} \geq 10$ ) ou (LABEL mod 16 = 1 et  $\text{pop75} < 10$ ) alors l'observation est une observation aberrante supplémentaire. Donc, le taux d'observations aberrantes est plus élevé pour les grandes municipalités. Cependant, il n'est pas influencé par les valeurs des autres variables. Nous pouvons dire que l'aberrance est aléatoire. Notons que nous aurions pu tirer un échantillon aléatoire au lieu de l'échantillon systématique susmentionné. Nous avons opté pour l'échantillon

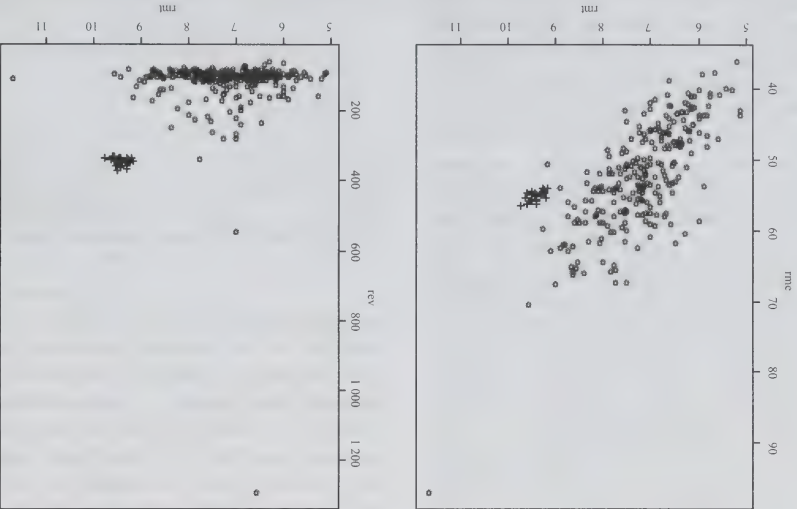
systématique afin de simplifier la répétition des résultats et d'éviter un caractère aléatoire supplémentaire.

Il se fait que deux des observations aberrantes de base pondérées et l'une de celles non pondérées sont aussi des observations aberrantes supplémentaires. Nous continuons de les traiter comme des observations aberrantes de base. Compte tenu de cette situation, il y a 32 observations aberrantes supplémentaires dans l'échantillon. Combienées aux 25 observations aberrantes de base non pondérées et aux 25 pondérées définies plus haut, elles donnent 57 ou 52 observations aberrantes (20,3 % ou 18,5 % des observations aberrantes). À partir d'ici, nous fixons le seuil pour les distances de Mahalanobis à la 57<sup>e</sup> distance la plus grande pour simplifier la comparaison des méthodes.

Les valeurs des observations aberrantes supplémentaires sont créées comme il suit :  $\text{rmt} = 0,2 * \text{rmt85} + 8$ ,  $\text{me} = 0,1 * \text{me84} + 50$ ,  $\text{rev} = 0,4 * \text{rev84} + 300$ . Notons que nous omettons le suffixe indiquant l'année pour les variables contaminées. La dépendance à l'égard des anciennes valeurs est négligeable. Il est utilisé uniquement pour éviter un modèle explicite pour l'erreur autour du point (rmt, me, rev) = (8, 50, 300). Il s'agit du type de contamination qu'il est difficile de détecter pour les estimateurs de covariance robustes (Rocke et Woodruff 1996), c'est-à-dire une contamination concentrée et proche du nuage de points des bonnes observations.

La figure 3 montre les trois variables contaminées et la position des observations aberrantes supplémentaires.

Figure 3 Variables me et rev c. rmt. Les observations aberrantes supplémentaires sont marquées d'une croix



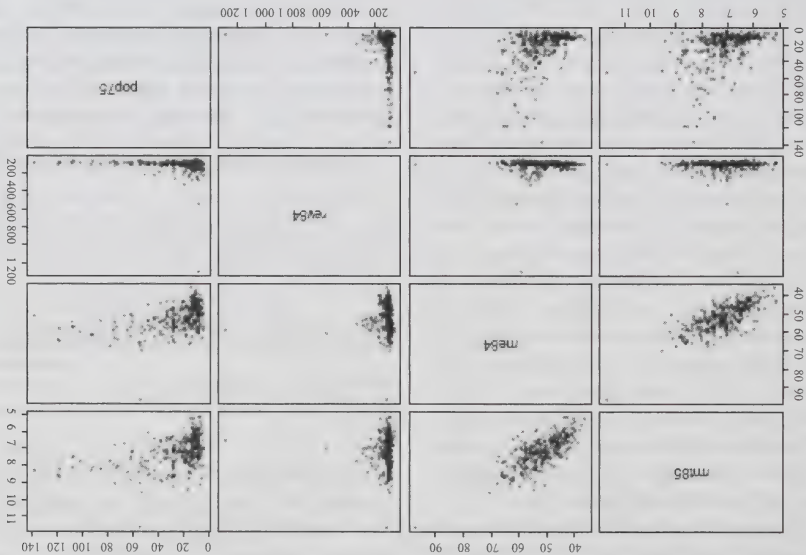


Figure 2 MU281 - Chiffres par habitant et pop75

### 5.2.1 Valeurs manquantes

Nous introduisons maintenant des valeurs manquantes dans les variables rm85, me84 et rev84 selon un mécanisme MAR léger. Nous attribuons un profil de valeurs manquantes aux observations en fonction du modulo de leur étiquette dans les données MU284 originales. Le mécanisme MAR est reflété par des taux plus élevés de valeurs manquantes pour les strates 1 et 2 (voir l'annexe). Les profils de réponse et les taux de valeurs manquantes par strate sont présentés au tableau 5. Par exemple, rev84 est la seule valeur manquante dans 15 observations de la strate 1 et dans deux observations de la strate 2. Dans l'ensemble, 187 observations demeurent complètes et la proportion d'observations pour lesquelles des valeurs manquent (taux de valeurs manquantes) est de 33 %.

Tableau 5 Fréquence des profils de réponse par strate pour rm85, me84 et rev84

Indicateur de réponse		Strate			
rm85	me84	rev84	Taux de valeurs manquantes		
			0,450	0,157	0,125
0	1	1	94	86	7
0	0	0	11	4	0
1	1	1	14	2	0
1	0	0	13	2	1
1	1	0	15	2	0
1	0	1	13	2	0
0	1	1	11	4	0
2	1	1	2	2	3

Parmi les 35 observations aberrantes de base pondérées ou non pondérées, il existe 17 observations incomplètes. Le tableau 6 indique combien de ces observations aberrantes de base ont été détectées après l'introduction des valeurs manquantes. Les 20 observations aberrantes de base pondérées sont bien détectées par les algorithmes pondérés ERW, BEMw et TRCW. GIMCD détecte 4 de ces observations et 14 des observations aberrantes de base non pondérées. Donc, la présence de valeurs manquantes affecte les capacités de l'algorithme MCD qui avait, en fait, été utilisé pour définir les observations aberrantes de base non pondérées. Une mise en garde est toutefois de rigueur. L'exécution répétée de l'imputation gaussienne a révélé une certaine variabilité des résultats de la méthode de GIMCD. Cependant, même avec un résultat d'imputation favorable, GIMCD ne bat pas les méthodes ER, BEM ou TRC dans la détection des observations aberrantes de base non pondérées. Les versions pondérées ERw, BEMw et TRCW détectent également les observations aberrantes de base pondérées. Le nombre d'observations complètes parmi les observations aberrantes désignées par les diverses méthodes est indiqué à la dernière ligne du tableau 6. Toutes les observations comportant des valeurs manquantes. Puisque le taux de valeurs manquantes est élevé dans la strate des petites observations et que les versions pondérées des méthodes désignent moins d'observations aberrantes dans cette strate, le nombre d'observations aberrantes complètes



d'échantillonnage ont été utilisés. Puisque les petites municipalités ont plus de poids, les estimations sont attirées vers elles et d'autres observations aberrantes apparaîtront parmi les 25 distances de Mahalanobis les plus grandes pour ERw, BEMw et TRCw (voir aussi le tableau 4). Une inspection plus approfondie révèle que le nombre des observations aberrantes non pondérées sont situées dans la queue de la distribution de rev84, tandis que la plupart des observations aberrantes pondérées sont situées dans la queue de la distribution de pop75. Les résultats des méthodes pondérées coïncident pour 20 observations considérées comme aberrantes. Nous les appellerons observations aberrantes de base pondérées. Ces dernières ont pour étiquette dans l'ensemble de données MU284 original 233, 241, 245, 265, 267 et 270. Seulement dix de ces observations figurent parmi les observations aberrantes de base non pondérées. Le tableau 4 donne le nombre d'observations aberrantes et non pondérées dans les strates. Dans la strate 1, il y a 12 observations aberrantes non pondérées, mais seulement deux pondérées. Donc, la pondération influe clairement sur la détection des observations aberrantes. L'influence s'exerce principalement sur le modèle qui est attiré vers les petites observations dont les poids sont grands. Naturellement, cette situation peut être considérée comme une sorte de masquage des observations aberrantes, mais dans le contexte de la modélisation, une meilleure explication est que le modèle n'est pas entièrement adéquat sur toutes les strates et que le modèle pondéré est mieux ajusté à la population que le modèle non pondéré.

La deuxième ligne du tableau 3 donne le temps de calcul des algorithmes. L'algorithme ER est nettement plus lent que ses concurrents, ce qui pourrait toutefois tenir à une implémentation inefficace. L'algorithme le plus rapide est BEM, suivi de TRC et, à une certaine distance, de MCD. L'algorithme TRC peut toutefois devenir lent si le taux de valeurs manquantes est élevé.

Tableau 3 MU281 complet, nombre d'observations aberrantes de base non pondérées détectées

Méthode	MCD	ER	ERw	BEM	BEMw	TRC	TRCw
Temps de calcul	0,81	3,17	2,52	0,07	0,04	0,14	0,14

Le suffixe w indique que l'algorithme est exécuté avec les poids d'échantillonnage.

Tableau 4 Nombre d'observations aberrantes de base par strate

Strate	1	2	3	Total
Non pondérées	12	5	8	25
Pondérées	2	10	8	20

Les trois variables RMT85, ME84 et REV84 sont divisées par pop85 pour obtenir les chiffres par habitant. Les variables dont la valeur est exprimée par habitant sont dénotées par des noms en lettres minuscules (mr85, me84 et rev84). La figure 2 montre la distribution de ces trois variables et de la variable auxiliaire pop75. Les chiffres par habitant suivent une distribution approximativement elliptique. Il y a un lien linéaire entre mr84 et me84 et un lien légèrement non linéaire entre ces variables et pop75. Il ne semble exister aucune relation entre rev84 et mr85, ni entre rev84 et me84, mais la variabilité de rev84 est clairement plus importante quand pop75 est faible. Les distributions de variables pop75 et rev84 sont asymétriques. Il existe une grande observation aberrante dans mr85 et dans me84, et au moins deux dans rev84. Nous incluons pop75 dans tous nos calculs. En pratique, on inclurait la variable auxiliaire qui définit le plan d'échantillonnage dans un modèle. Notons que pop75 ne présente pas de valeurs manquantes.

Le diagramme qq des distances de Mahalanobis basé sur MCD montre uniquement les deux observations aberrantes manifestées dans rev84. La grande observation aberrante relevée conjointement dans mr85 et dans me84 possède la 25<sup>e</sup> plus grande distance de Mahalanobis. Nous donnons aux 25 observations les plus grandes le nom d'observations aberrantes de base non pondérées. Dans l'ensemble de données MU284 original, ces observations ont les étiquettes 3, 4, 29, 31, 46, 79, 83, 117, 126, 131, 140, 158, 199, 211, 222, 246, 248, 252, 254, 260, 262, 272 et 273. Dans le cas des distances de Mahalanobis non robustes classiques, nous fixons le nombre d'observations qui doivent être considérées comme aberrantes à 25 pour le moment. Donc, nous considérons pour chaque méthode les 25 observations pour lesquelles la distance de Mahalanobis est la plus grande comme étant des observations aberrantes. Soulignons qu'il ne s'agit pas forcément du seuil que l'on choisirait après inspection du graphique qq des distances de Mahalanobis.

Le tableau 3 montre le nombre d'observations aberrantes de base détectées par les méthodes exécutées sur l'ensemble de données complètes. Évidemment, MCD détecte ses propres 25 observations aberrantes. Les algorithmes ER, BEM et TRC détectent 25 ou 24 de ces observations aberrantes si aucune pondération n'est appliquée, mais uniquement 11 ou 15, s'ils sont pondérés. Un suffixe « w » derrière l'acronyme de la méthode indique que les poids

TRC détecte 13 observations aberrantes et déclare qu'une bonne observation est aberrante.

## 5.2 Données MU281

L'ensemble de données MU284 tiré de Särnadal, Swensson et Wretman (1992) contient des données sur les municipalités suédoises. Nous utilisons les variables *population en 1975* et *population en 1985* (pop75 et pop85), *recettes des impôts municipaux 1985* (RMT85), *nombre d'employés municipaux 1984* (ME84) et *valeur immobilière 1984* (REPV84). Les trois plus grandes villes selon pop75 sont écartées, parce qu'elles représentent d'énormes observations aberrantes qui seraient traitées séparément en pratique. Nous supposons que les municipalités restantes constituent un échantillon stratifié d'une population plus grande. Les strates sont définies conformément à  $0 < \text{pop75} < 20$ ,  $20 \leq \text{pop75} < 100$ ,  $100 \leq \text{pop75}$ . Le tableau 2 donne les tailles de population supposées et les poids correspondants. Ce plan d'échantillonnage reflète une stratification typique des enquêtes auprès des établissements avec une strate à tirage complet pour les établissements plus grands, où, en bout de ligne, huit des dix établissements répondent à l'enquête.

Tableau 2 Taille de population et d'échantillon pour MU281

strate	1	2	3
pop75	0 à 19	20 à 99	100+
N	1 600	250	10
n	171	102	8
w	9,36	2,45	1,25

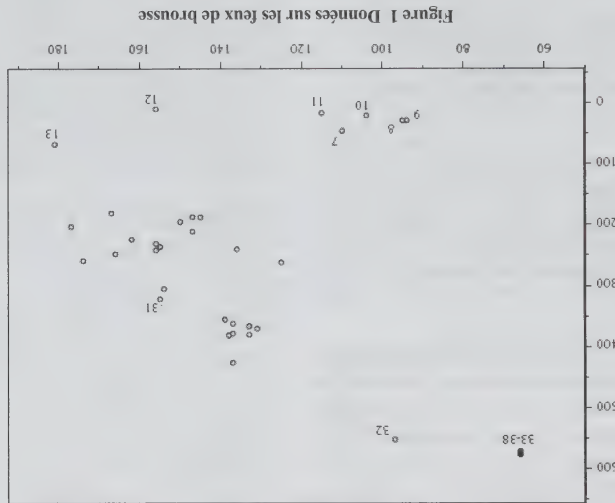


Figure 1 Données sur les feux de brousse

Dans le cas où 40 % de valeurs manquent, l'algorithme ER désigne les observations 7 et 13 et deux bonnes observations comme étant aberrantes. Puisque l'imputation est aléatoire, le résultat de GIMCD présente une certaine variabilité. Deux réalisations, GIMCD(1) et GIMCD(2) sont présentées au tableau 1. Dans un bon cas, GIMCD détecte 10 des 15 observations aberrantes et ne désigne aucune bonne observation comme étant aberrante. Dans un mauvais cas, GIMCD ne détecte que quatre observations aberrantes, mais indique que cinq bonnes observations sont aberrantes. La méthode BEM détecte 12 des observations aberrantes et déclare qu'une bonne observation est aberrante. La méthode

Dans le cas où 20 % des valeurs manquent, les méthodes BEM et TRC désignent un bon point de donnée comme étant une observation aberrante. La méthode TRC détecte 14 des 15 observations éventuellement aberrantes. La méthode ER manque toutes les observations aberrantes, sauf l'observation 7. La méthode GIMCD détecte les 13 mêmes observations aberrantes que dans la situation où aucune valeur ne manque. Notons toutefois qu'il existe une certaine variabilité dans les résultats de la GIMCD à cause de l'imputation aléatoire. La méthode BEM détecte les mêmes observations aberrantes que sous les conditions de données complètes, sauf l'observation 11.

Dans le cas où 20 % des valeurs manquent, les méthodes BEM et TRC désignent un bon point de donnée comme étant une observation aberrante. La méthode TRC détecte 14 des 15 observations éventuellement aberrantes. La méthode ER manque toutes les observations aberrantes, sauf l'observation 7. La méthode GIMCD détecte les 13 mêmes observations aberrantes que dans la situation où aucune

Béguin et Hulliger : L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées

3. Calculer les carrés des distances de Mahalanobis marginales  $MD_G^2(x_i)$  pour  $i=1, \dots, n$ . Le nouvel ensemble  $G'$  contient les observations pour lesquelles  $MD_G^2(x_i) < c^{N_{p.d.}}$
4. Si  $G' = G$ , arrêter; sinon, fixer  $G \leftarrow G'$  et passer à l'étape 2.

Si les objectifs principaux de l'algorithme BACON-EEM sont les estimations de la moyenne et de la covariance plutôt que la détection des observations aberrantes, l'algorithme EEM peut être plus loin sans changer  $G$ . À l'étape 3, on pourrait aussi utiliser  $\alpha/n$  au lieu de  $\alpha$  (voir la section 2).

## 5. Applications

À la présente section, nous comparons l'algorithme BACON-EEM (BEM) à la méthode des corrélations de rang transformées (TRC) de Bégin et Hurliger (2004) et à l'algorithme ER de Little et Smith (1987). En guise de valeurs repères supplémentaires, nous utilisons une applica-

En effet, nous créons une imputation non robuste. Puis, moyennant l'algorithme EM, nous obtenons des estimations robustes de la position multivariée et de la matrice de covariance à l'aide de l'estimateur par déterminant minimal de covariance (MCD) calculé sur les données imputées et, enfin, nous détectons les observations aberrantes en utilisant les distances de Mahalanobis correspondantes. La méthode de référence est appelée *GMCD pour Gauss imputation followed by MCD detection* (imputation gaussienne suivie d'une détection MCD). Les algorithmes sont implémentés en R (R Development Core Team 2006) avec l'aide des packages *R norm* (Novo et al 2002) et *MASS* (Venables et Ripley 2002).

## 5.1 Données sur les feux de brousse

La réaction de l'algorithme BACON-EEM à l'intro-  
duction de valeurs manquantes est illustrée à l'aide d'un  
ensemble de données réelles de 38 observations et  
5 variables. Il a été utilisé par Maronna et Zamar (2002)  
pour localiser les empreintes de feux de brousse. Cet  
exemple bien connu est également étudié dans Maronna et  
Yohai (1995) et dans Maronna et Zamar (2002). Il permet  
de produire un graphique bidimensionnel (dans les  
variables 2 et 3) qui révèle la plupart des observations  
aberrantes (voir la figure 1). Les données contiennent une  
groupe d'observations aberrantes 33 à 38, une deuxième  
groupe d'observations aberrantes 7 à 11, et quelques autres  
observations aberrantes isolées, à savoir les observations 12,  
13, 31 et 32. Nous avons ajouté l'observation 31 à la liste  
des observations aberrantes éventuelles, parce que les  
méthodes MCD, BACON et d'autres méthodes étudient

Tableau 1  
Observations aberrantes détectées pour trois taux  
de valeurs manquantes

(1 - g) % Méthode	7-11	12,13,21	32-38	N <sup>bre</sup>
0	11100	000	0000000	0
0	11111	001	1111111	0
0	11111	100	0000000	0
0	11111	001	1111111	0
0	11111	000	1111111	0
0	01100	000	1111111	0
20	11111	001	1111111	0
20	10000	000	0000000	0
20	ER			0
20	GIMCD			0
40	GIMCD(1)			0
40	11100	000	1111111	0
40	11100	000	0100000	5
40	ER	010	0000000	2
40	BEM	000	1111111	1
40	TRC	11111	1111111	1

dans Maronna et Zamara (2002) indiquent qu'il s'agit d'un cas limite. Les valeurs manquantes sont créées selon un mécanisme MCAR (valeur manquant entièrement au hasard). Deux ensembles de données sont créés en utilisant respectivement 20 % et 40 % d'items manquants. L'ensemble de données pour lequel 40 % des valeurs manquantes contient des observations pour lesquelles jusqu'à quatre valeurs sur cinq manquent et, par conséquent, présentent un défi quelle que soit la méthode. Comme la taille  $n$  de l'ensemble de données est faible, nous exécutons l'algorithme BACON-EEM avec le test  $\chi^2_{dof,n}$ . Les résultats sont présentés au tableau 1. Pour les observations 7 à 13 et 31 à 38, nous indiquons individuellement si elles sont détectées ou non, tandis que pour les 23 autres bons points, nous indiquons le nombre d'observations désignées comme étant aberrantes. La limite au dessous de laquelle une distance de Mahalanobis indique qu'une valeur est aberrante a été déterminée pour chaque passage machine en examinant le graphique des quantiles de la distance de Mahalanobis.

MD : distance de Mahalanobis classique, MCD : déterminant minimal de covariance, GIMCD : imputation non robuste sous modèle gaussien suivi de MCD (GIMCD1) et GIMCD(2) (sans deux réalisations de l'algorithme GIMCD), ER : espérance-maximisation avec une étape M à la maximisation, BEM : BACON-EM, TRC : la proportion de valeurs manquantes, la dernière colonne indique les corrélations de rang transformées. La première colonne indique le nombre d'autres points (non aberrants) déclarés comme des observations aberrantes, les colonnes intermédiaires sont des indicateurs de détection des observations dans la première ligne.

(Quand les observations sont complètes, aucune observation n'est désignée comme étant aberrante par aucune des méthodes. La distance de Mahalanobis avec moyenne et covariance non robustes détecte trois observations aberrantes, mais manquent les autres. Les méthodes MCD et BACON-EBM aboutissent à la sélection du même sous-ensemble de données comme sous-ensemble de bons points et, par conséquent, donnent le même résultat, c'est-à-dire exactement la même distance de Mahalanobis pour chacune



aberrante. Par conséquent, l'imputation avant la détection des observations aberrantes n'offre aucun avantage et nous préférons adapter directement la distance de Mahalanobis à la situation où des valeurs manquent. Deux versions distinctes de cette distance sont possibles dans cette situation.

Nous appelons la première version distance de Mahalanobis *marginale*. Elle consiste à utiliser la distance de Mahalanobis dans l'espace des variables observées et à la rééchantillonner à l'aide d'un facteur  $p/q$ , où  $q = \sum_k r_k$  est le nombre de variables non manquantes et  $p$  est le nombre total de variables. Plus précisément, nous supposons qu'une observation  $x$  est partitionnée en  $x = (x_0^T, x_m^T)^T$  (éventuellement après un réarrangement), où  $x_0$  dénote la partie observée et  $x_m$  la partie non observée de l'observation. Alors, la distance de Mahalanobis marginale est

$$MD^2_{\text{marg}} = \frac{b}{d} (x_0 - m_0)^T (S_{00})^{-1} (x_0 - m_0), \quad (9)$$

où  $S_{00}$  est la partie de la matrice de covariance qui correspond à  $x_0$ . Cette version est également utilisée dans Little et Smith (1987).

La deuxième version de la distance de Mahalanobis en cas de valeurs manquantes s'obtient en réduisant à zéro la contribution des valeurs manquantes à cette distance. Cela reviendrait à remplacer toutes les valeurs manquantes par leur moyenne, c'est-à-dire  $x_m = m_m$ . Autrement dit, nous imputerions une moyenne sans tenir compte de la matrice de covariance et les arguments à l'encontre de la détection des observations aberrantes avec valeurs imputées s'appliqueraient également ici. Néanmoins, nous avons testé cette deuxième version de la distance de Mahalanobis. Elle produit des distances erratiques (Bégin 2002 ainsi que Bégin et Hurliger 2003) et nous n'avons pas poursuivi son utilisation.

#### 4. L'algorithme BACON-EEM

Les algorithmes BACON et EEM demandent tous deux une grande puissance de calcul et en les fusionnant d'une manière commode, nous augmentons la performance. La structure de « croissance » de l'algorithme BACON implique des redondances qui peuvent être utilisées pour éviter des calculs supplémentaires à chaque étape de l'algorithme EEM. À chaque étape BACON, le point crucial est que les estimations de la moyenne et de la matrice de covariance à partir de l'algorithme EEM permettent d'exclure les points aberrants du bon sous-ensemble, ce qui ne requiert pas d'estimations extrêmement précises. Donc, il n'est pas nécessaire d'itérer EEM jusqu'à la convergence chaque fois que la moyenne et la covariance sont requises. Nous utilisons seulement cinq

itérations par défaut. En outre, nous prenons le résultat de la dernière itération EEM de la dernière étape BACON comme valeur de départ pour l'algorithme EEM. Avant d'information que possible provenant des itérations antérieures devrait être réutilisée. En fait, la statistique suffisante  $T^c$  calculée sur un bon sous-ensemble  $G$  possède une partie observée de la somme  $T^c_0$  et une partie manquante de la somme  $T^c_m$ . L'espérance calculée par l'étape E peut par conséquent s'écrire

$$E(T^c | X^c_0, \theta) = T^c_0 + E(T^c_m | X^c_0, \theta). \quad (10)$$

Comme les sous-ensembles  $G$  grandissent habituellement,  $T^c_0$  n'est pas recalculé à chaque étape de la boucle BACON, mais une variable globale pour  $T^c_0$  est mise à jour chaque fois que  $G$  change (habituellement l'ajout de points seulement, parfois l'élimination de quelques-uns). À chaque itération de l'algorithme BACON-EEM, quand l'algorithme EEM a obtenu les estimations du centre et de la dispersion du bon sous-ensemble, les distances de Mahalanobis marginales de toutes les observations sont utilisées à l'étape 2 de l'algorithme BACON.

Mentionnons le point crucial en ce qui concerne la robustesse de l'algorithme : EEM n'est pas robuste, mais à chaque étape BACON, il est exécuté uniquement sur les points dont la distance de Mahalanobis marginale était la plus petite et, par conséquent, non aberrante à l'étape précédente. Autrement dit, l'observation  $x$  sera utilisée par EEM si, et uniquement si,  $x_0$  est suffisamment petit pour la mesure donnée par  $(S_{00})^{-1}$  à l'étape précédente. Par conséquent, si le premier sous-ensemble de bons points est exempt ou presque exempt d'observations aberrantes, le processus d'imputation dans EEM ne créera jamais d'observations aberrantes au cours de l'exécution complète de l'algorithme BACON-EEM. Cela signifie que l'algorithme EEM non robuste est protégé par l'approche de recherche avant générale de l'algorithme BACON, de la même façon que sont protégées la moyenne et la covariance non robustes de l'algorithme BACON original.

En résumé, les étapes de l'algorithme BACON-EEM sont les suivantes :

1. Calculer la médiane des coordonnées pondérée  $\text{med}(x)$  en ignorant les valeurs manquantes dans chaque variable séparément. Déterminer la distance euclidienne par rapport à la médiane de chaque observation en omettant les valeurs manquantes, mais en normalisant pour le nombre de valeurs présentes :  $a_i = \|x_i - \text{med}(x)\| \sqrt{d/q}$ . Sélectionner les  $m = cp$  observations ayant les  $a_i$  les plus faibles pour constituer le sous-ensemble  $G$  initial.
2. Calculer un centre  $m_G$  et une dispersion  $C_G$  en utilisant l'algorithme EEM et mettre à jour l'estimation de la statistique suffisante de la partie observée  $T^c_0$ .

sous les données observées sont décrites dans Dempster et coll. (1977). Dans le cas de problèmes « dociles », ce point stationnaire est un maximum global.

Pour une loi de probabilité appartenant à la famille exponentielle ordinaire, la log-vraisemblance sous données complètes peut s'écrire

$$l(\theta | X) = \eta(\theta)^\top \cdot T(X) + Ng(\theta) + c, \quad (4)$$

où  $\eta(\theta) = (\eta_1(\theta), \eta_2(\theta), \dots, \eta_q(\theta))^\top$  est la forme canonique du paramètre  $\theta$  et  $T(X) = (T_1(X), T_2(X), \dots, T_r(X))^\top$  est le vecteur des statistiques suffisantes sous données complètes. En outre, chacune des statistiques suffisantes possède une forme additive  $T_j(X) = \sum_{i=1}^N h_j(x_i)$  pour une certaine fonction  $h_j$ . Comme  $l(\theta | X)$  est une fonction linéaire des statistiques suffisantes, dans l'étape E,  $T_j(X)$  est remplacé par  $E(T_j(X) | X_{\circ}, \theta_{(i)})$ . Autrement dit, l'étape E comble les parties manquantes des statistiques suffisantes sous données complètes. Pour une loi normale multivariée  $X = (X_1, \dots, X_p)$ , les statistiques suffisantes sont composées de deux types d'éléments : les sommes  $\sum_{i=1}^N x_i^l x_i^k$  et les sommes des produits  $\sum_{i=1}^N x_i^l x_i^k x_i^l$ ,  $1 \leq k, l \leq p$ . L'étape E se réduit au calcul des espérances conditionnelles de ces sommes sachant les données observées  $X_{\circ}$  et le paramètre provisoire  $\theta_{(i)}$ .

Pour un seul indice de sommation  $i$ , on peut montrer (Schafer 2000, section 5.3) que ces espérances dépendent uniquement des composantes observées de la même observation, c'est-à-dire de  $x_{i, \text{obs}}$ . Cela nous mène à

$$E \left( \sum_{i=1}^N x_i^l x_i^k \middle| X_{\circ}, \theta_{(i)} \right) = \left( \sum_{i=1}^N E(x_i^l x_i^k | X_{\circ}, \theta_{(i)}) \right) = \sum_N E(x_i^l x_i^k | x_{i, \text{obs}}^l, \theta_{(i)}), \quad 1 \leq k \leq p \quad (5)$$

et à la forme analogue de la somme des produits. Naturellement,  $E(x_i^k | x_{i, \text{obs}}^l, \theta_{(i)}) = x_i^k$  si  $x_i^k \in x_{i, \text{obs}}^l$ . Si  $x_i^k$  manque, alors cette espérance est la valeur ajustée d'une régression de  $x_i^k$  sachant le paramètre  $\theta_{(i)}$  sur les variables qui sont observées pour l'observation  $i$ . Donc, les statistiques suffisantes sont composées des sommes de population des valeurs observées ( $T_o$ ) et des sommes des valeurs ajustées ( $T_m$ ).

Dans la situation où nos données proviennent d'un échantillon d'une population finie, nous considérons la population finie comme une réalisation de la loi normale multivariée, et les sommes ( $\Sigma$ ) et les sommes des produits doivent être estimées d'après l'échantillon. La forme de (5) nous permet d'utiliser les simples estimateurs d'Horvitz-Thompson. L'estimation de (5) est

$$T_{k=0} = \sum w_i E(x_i^k | x_{i, \text{obs}}^l, \theta_{(i)}), \quad 1 \leq k \leq p, \quad (6)$$

et  $E(\sum_{i=1}^N x_i^l x_i^k | X_{\circ}, \theta_{(i)})$  est estimée par

$$T_{kl} = \sum w_i E(x_i^l x_i^k | x_{i, \text{obs}}^l, \theta_{(i)}), \quad 1 \leq k, l \leq p. \quad (7)$$

Brièvement, nous remplaçons les sommes de population de  $T_o$  et  $T_m$  par les estimateurs d'Horvitz-Thompson  $\hat{T}_o$  et  $\hat{T}_m$ . Nous appelons le calcul de  $T_{k=0}$  et de  $T_{kl}$  l'étape de l'espérance estimée (étape EB). En introduisant ces estimateurs dans l'expression (4), nous obtenons un estimateur de la fonction de vraisemblance de population moyenne.

Pour l'étape M, c'est-à-dire la maximisation de l'estimation de la vraisemblance de population moyenne, nous devons résoudre les équations normales pondérées. Nous obtenons la solution par une simple opération matricielle en appliquant l'opérateur SWEEP (Schafer 2000, section 5.3) à la matrice symétrique  $(T_{kl})_{0 \leq k, l \leq p}$  dimensions  $(p+1) \times (p+1)$  des espérances estimées des statistiques suffisantes (avec  $T_{00}$  fixé à 1) divisée par  $N$ , qui est estimé par la somme des poids s'il est inconnu :

$$\theta_{(i+1)} = \text{SWP}[0] \left( \frac{\sum_{0 \leq k, l \leq p} w_i}{d} \right), \quad (8)$$

où SWP[0] est l'opérateur SWEEP sur la première ligne/colonne de la matrice.

L'algorithme EEM itère les étapes EB et M. Du point de vue des calculs, la différence entre l'étape EB et l'étape E de l'algorithme EM original tient à l'utilisation de sommes pondérées au lieu de sommes non pondérées en ne faisant pas varier les poids au cours des itérations. Par conséquent, nous nous attendons à ce que la convergence de l'algorithme EEM reste semblable à celle de l'algorithme EM. Pour l'algorithme BACON-EEM, nous avons seulement besoin d'une approximation grossière de la solution à chaque étape BACON. Donc, nous n'utilisons qu'un petit nombre d'itérations de l'algorithme EEM.

### 3.1 Distance de Mahalanobis avec valeurs manquantes

La distance de Mahalanobis est établie pour des observations complètes et doit être adaptée aux observations incomplètes. Une option consiste à utiliser l'estimation EEM pour imputer la moyenne conditionnelle pour les valeurs manquantes, sachant les valeurs observées, puis à calculer la distance de Mahalanobis en utilisant les valeurs imputées. Sous une hypothèse MAR (données manquant au hasard), il existe un modèle valide fondé sur la partie observée des données pour imputer les valeurs manquantes. Dans le cas de la détection des observations aberrantes, nous supposons que le modèle d'imputation peut tenir pour la majeure partie des données seulement et qu'il est estimé de manière robuste. Mais alors, nous ne pouvons pas nous attendre à ce qu'une valeur aberrante soit prédite par le modèle, sauf si la partie observée d'une observation est déjà

Afin d'adapter l'algorithme BACON à l'échantillonnage, nous sélectionnons le sous-ensemble initial en utilisant les estimateurs de Hajék  $m_X$  et  $C_X$  ou la médiane  $\text{med}_X$ . Pour le processus itératif, nous dénotons par  $s_G$  les « bonnes observations » sélectionnées dans l'échantillon. Ces observations sont représentatives d'un « bon sous-ensemble virtuel »  $G$  de l'ensemble de la population dont la taille estimée est  $F = \sum_{s_G} w_i$ . La moyenne et la matrice de covariance de ce sous-ensemble sont estimées par les estimateurs de Hajék

$$\hat{m}_G = \frac{\sum_{s_G} w_i x_i}{\sum_{s_G} w_i}$$

(3)

$$\hat{C}_G = \frac{\sum_{s_G} w_i (x_i - \hat{m}_G)(x_i - \hat{m}_G)^T}{\sum_{s_G} w_i}$$

Nous utilisons ces estimations pour calculer les estimations des distances de Mahalanobis  $\text{MD}_G(x_j)$ ,  $x_j \in s$ . Enfin, nous calculons le facteur de correction  $c_j^{\text{Ndp}} = c_j^{\text{Ndp}} + c_j^{\text{Hr}}$  du critère de sélection en utilisant les estimations  $\hat{N} = \sum_{s_G} w_i$  et  $F = \sum_{s_G} w_i$ . Si  $N$  est connu, nous utilisons sa valeur réelle.

S'il n'y a pas de valeurs manquantes dans les données, l'algorithme BACON peut être utilisé pour estimer la moyenne et la covariance de population. L'hypothèse fondamentale pour l'algorithme BACON est encore que la majorité des observations de la population suivent une loi elliptique. Nous pouvons utiliser l'algorithme BACON sans pondération et comparer les résultats à la version pondérée. L'obtention de résultats différents indique que les variables du plan ou le modèle utilisé pour la pondération de la non-réponse ne sont pas bien reflétés dans le modèle. Nous recommandons d'utiliser les poids, en particulier dans les applications de routine, afin d'offrir une certaine protection contre l'erreur de spécification du modèle. Quoiqu'il en soit, les paramètres estimés devraient être la moyenne et la covariance de la majorité de la population.

Souignons que la distance de Mahalanobis ne fait pas intervenir directement les poids d'échantillonnage. Le poids d'une observation éventuellement aberrante n'influe sur la distance de Mahalanobis que par la voie du modèle, c'est-à-dire la moyenne et la covariance. Les problèmes de non-réponse sont importants en statistiques officielle et de nombreux sondages ne peuvent pas fournir un ensemble de données complètes. Le problème de la non-réponse totale, c'est-à-dire des observations dont toutes les valeurs manquent, est habituellement résolu par utilisation de pondérations appropriées et il n'est

### 3. L'algorithme EEM

Nous réutilisons dans ce paragraphe la description et la notation de l'algorithme EM donnée dans Schaffer (2000). Tous les détails au sujet de l'algorithme non présentés ici peuvent être consultés dans les trois premiers chapitres et à la section 5.3 de ce livre. Nous recourons également ici à l'abus de notation suivant : le  $X$  dénotera simultanément une variable aléatoire de dimension  $p$  et la matrice de dimension  $N \times p$  contenant les valeurs réalisées de la variable  $X$  de la population  $U$ . Si l'on procédait à un recensement de l'ensemble de la population pour mesurer la variable  $X$ , on obtiendrait un certain ensemble de valeurs observées et de valeurs manquantes  $X = X_o \cup X_m$ . L'algorithme EM repose sur l'hypothèse que le mécanisme de création des données manquantes est ignorable (Schaffer 2000 section 2.2). Ici, nous supposons en outre que la présence de données manquantes est indépendante de l'échantillonnage. Les observations des tirages indépendants être modélisées comme étant des tirages indépendants et identiquement distribués (iid) à partir d'une loi de probabilité normale multivariée de densité  $f(x, \theta)$ . En utilisant les hypothèses et la factorisation  $P(X|\theta) = P(X_o|\theta)P(X_m|X_o, \theta)$  nous pouvons écrire la log-vraisemblance sous données complètes sous la forme  $l(\theta|X) = l(\theta|X_o) + \log(P(X_m|X_o, \theta)) + c$ , où  $l(\theta|X_o)$  est une constante choisie arbitrairement. Le terme  $P(X_m|X_o, \theta)$  reflète l'interdépendance de  $X_m$  et de  $\theta$  dont tire part l'algorithme EM. Comme  $P(X_m|X_o, \theta)$  est inconnue, la moyenne de  $l(\theta|X)$  sur  $P(X_m|X_o, \theta^{(t)})$  est prise à chaque étape  $E$ , où  $\theta^{(t)}$  est une estimation provisoire du paramètre inconnu. L'estimation suivante  $\theta^{(t+1)}$  est trouvée en maximisant le résultat de l'étape de l'espérance (étape  $M$ ). La séquence d'étapes  $E$  et  $M$  est itérée jusqu'à la convergence. Les conditions sous lesquelles cette séquence  $\theta^{(t)}$  converge vers un point stationnaire de la vraisemblance



$x_1^1, \dots, x_p^1$ . Nous supposons que la majorité d'entre-elles sont unimodales et approximativement elliptiques symétriques. La moyenne des coordonnées (matrice de covariance, resp.) calculée sur  $X$  est dénotée par  $m_X$  ( $C_X$ , resp.). Le carré de la distance de Mahalanobis d'un point  $p$  basé sur  $m_X$  et  $C_X$  est  $MD_X^2(p) = (p - m_X)^T C_X^{-1} (p - m_X)$ . Si la moyenne et la covariance sont calculées uniquement sur un sous-ensemble  $G$  des données, nous les dénotons  $m_G$  et  $C_G$  avec la distance de Mahalanobis correspondante  $MD_G$ .

La première étape de l'algorithme est le choix d'un sous-ensemble initial  $G$  de « bonnes données ». Deux versions sont proposées dans la littérature. La première consiste simplement à sélectionner les  $cp$  points ayant les plus petites distances de Mahalanobis  $MD_X(x_i)$ ,  $i \in \{1, \dots, n\}$ , où  $c$  est un entier choisi par l'analyste des données. Sa valeur peut être fixée à  $c = 3$  par défaut. La deuxième version consiste à choisir les  $cp$  points ayant les plus petites distances euclidiennes par rapport à la médiane des coordonnées, avec  $c$  défini comme auparavant. La variance affine. D'autres points de départ que la médiane des coordonnées pourraient être envisagés, comme une médiane spatiale. Dans cet article, nous nous concentrons sur la deuxième version du bon sous-ensemble de base. Dans les deux versions, si  $C_G$  est singulière, le sous-ensemble de base est étendu en ajoutant les observations ayant les plus petites distances jusqu'à ce que  $C_G$  soit de plein rang. Après cela, le processus itératif commence.

Soit  $\chi_{1-\beta}^2$  le  $1 - \beta$  centile de la loi  $\chi^2$  avec  $p$  degrés de liberté et  $|G|$ , le nombre d'éléments dans l'ensemble  $G$ . Les étapes de l'algorithme BACON sont :

1. Calculer les carrés des distances de Mahalanobis  $MD_G^2(x_i)$  pour  $i \in \{1, \dots, n\}$ .
2. Définir un sous-ensemble  $G'$  incluant tous les points pour lesquels  $MD_G^2(x_i) < c_{np}^{np, c/m}$  où  $c_{np}^{np, c/m} = c_{np}^{np} + c_{np}^{np}$  est un facteur de correction avec  $c_{np}^{np} = 1 + (p + 1)/(n - p - d) + 1/(n - p - d)$ ,  $c_{np}^{np} = \max\{0, (h - p)/(h + p)\}$ ,  $h = \lceil n + p + 1/2 \rceil$  et  $r = |G|$ .
3. Si  $G' = G$ , arrêter; sinon, fixer  $G$  à  $G'$  et aller à l'étape 1.

Souignons que le facteur de correction  $c_{np}^{np}$  est proche de 1 si  $n$  est grand. Les observations qui ne sont pas contenues dans l'ensemble de données  $G$  final sont déclarées aberrantes. Une autre approche consiste à choisir pour la distance de Mahalanobis un seuil au-delà duquel les observations sont déclarées aberrantes en inspectant la distribution de la distance de Mahalanobis. La complexité des calculs que requiert l'algorithme BACON dépend de la configuration des données. Comparativement à d'autres algorithmes, elle est faible et, en

particulier, augmente lentement avec la taille de l'échantillon (voir aussi la section 5). La méthode BACON convient donc particulièrement bien pour les grands ensembles de données.

Notons que le critère de sélection original de l'étape 2 est conçu pour une loi normale multivariée, qui implique que les carrés des distances de Mahalanobis suivent asymptotiquement une loi  $\chi^2$  avec  $p$  degrés de liberté. Supposons que tous les points suivent une loi normale multivariée et que la distance de Mahalanobis est calculée en utilisant la moyenne d'échantillon et la matrice de covariance. Le test  $MD_X^2(x_i) > \chi_{1-\alpha}^2$  déclare qu'environ 100 $\alpha$  pour cent des points sont des observations aberrantes. Au lieu d' $\alpha$ , nous utilisons souvent  $\alpha/n$ . À l'aide des inégalités de Bonferroni, on peut montrer que, sous la loi normale, le test avec le niveau  $\alpha/n$  n'aboutit à la déclaration d'aucune observation aberrante ayant une probabilité supérieure à  $1 - \alpha$  (c'est-à-dire  $P(MD_X^2(x_i) < \chi_{1-\alpha/n}^2 \mid \forall i \in \{1, \dots, n\}) \geq 1 - \alpha$ ). L'utilisation de  $\alpha/n$  pour le test même très rarement à la détection de points qui ne sont pas des observations aberrantes, mais elle réduit aussi la sensibilité aux observations aberrantes proches quand  $n$  devient grand. Il pourrait être utile d'exécuter la méthode avec les deux types de niveau de test et de comparer les résultats.

## 2.1 Adaptation aux poids d'échantillonnage

Dans le contexte de l'échantillonnage, nous utilisons la notation qui suit. Les données proviennent d'un échantillon aléatoire  $s$  de la population finie  $U$  contenant les  $N$  éléments. L'échantillon de taille  $n$  est tiré selon le plan d'échantillonnage  $p(s)$  et les probabilités d'inclusion de premier ordre sont dénotées  $\pi_i = \sum_{s \in U} p(s)$ . Les poids seront les inverses des probabilités d'inclusion des observations  $w_i = 1/\pi_i$ , tels que l'estimateur d'Horvitz-Thompson du total de population,  $\sum_{i \in U} x_i$ , soit  $\sum_s w_i x_i$ , soit  $\sum_{i=1}^n w_i x_i$ . En outre, nous supposons que  $\sum_s w_i \approx N$ . La moyenne  $m_X$  et la matrice de covariance  $C_X$  peuvent être estimées par les estimateurs de Hajek

$$\hat{m}_X = \frac{\sum_s w_i x_i}{\sum_s w_i}, \quad \hat{C}_X = \frac{\sum_s w_i (x_i - \hat{m}_X)(x_i - \hat{m}_X)^T}{\sum_s w_i} \quad (1)$$

L'estimation d'échantillon de la médiane est définie comme dans Béguin et Hulliger (2004) : soit  $x_k^n$  la valeur la plus petite telle que  $\sum_s w_i 1_{x \leq x_k^n}(x_i) \geq 0,5 \sum_s w_i$  et  $x_k^n$  la valeur la plus petite telle que  $\sum_s w_i 1_{x \leq x_k^n}(x_i) > 0,5 \sum_s w_i$  alors, l'estimation est donnée par

$$\widehat{\text{med}}_X = (w_k^n x_k^n + w_{k+1}^n x_{k+1}^n) / (w_k^n + w_{k+1}^n). \quad (2)$$

classiques et, par conséquent, son point de défaillance est 0. Autrement dit, même une seule observation aberrante peut porter l'estimateur au-delà de toute limite. Pour corriger le faible point de défaillance de cet algorithme, Cheng et Victoria-Feser (2000) ont utilisé un algorithme MCD pour l'étape de maximisation de l'algorithme EM. Cependant, la combinaison des procédures itératives du MCD et de l'EM rend les calculs trop lents pour les applications pratiques quand les ensembles de données sont grands. En outre, l'introduction des poids d'échantillonnage n'est pas simple. L'algorithme TRC de Béguin et Hülliger (2004) utilise des imputations par régression linéaire robuste par le meilleur prédicteur univarié pour traiter les valeurs manquantes. Les corrélations de rang de Spearman sont exprimées comme des fonctionnelles de la fonction de répartition empirique de l'échantillon pour obtenir des estimations des corrélations de rang de Spearman dans la population.

Comme l'algorithme BACON est fondé sur la loi normale multivariée, nous avons choisi l'algorithme EM pour données normales multivariées pour imputer les valeurs manquantes dans les itérations BACON. Afin de tenir compte de l'échantillonnage, nous avons remplacé les estimations de l'algorithme BACON par des estimateurs de type Horvitz-Thompson et nous avons élaboré une version spéciale de l'algorithme EM où les espérances au niveau de la population sont estimées d'après l'échantillon. À la section 2, nous décrivons la notation, rappelons rapidement en quoi consiste l'algorithme BACON et présentons son adaptation aux poids d'échantillonnage. À la section 3, nous introduisons l'algorithme d'espérance estimée/maximisation (EEM) et à la section 1, nous discutons de l'adaptation de la distance de Mahalanobis aux valeurs manquantes. À la section 4, nous expliquons comment les algorithmes BACON et EEM sont fusionnés de manière efficace pour devenir l'algorithme BACON-EEM. À la section 5, nous montrons l'application de l'algorithme BACON-EEM à deux ensembles de données. Nous comparons les résultats aux méthodes concurrentes, c'est-à-dire la méthode des corrélations de rang transformées (TRC) élaborées dans Béguin et Hülliger (2004), l'algorithme ER et un algorithme fondé sur l'algorithme EM.

## 2. L'algorithme BACON

L'algorithme BACON est présenté dans Billor et coll. (2000). Deux versions sont décrites : l'une pour les données multivariées en général et l'autre pour les données de régression. Nous ne considérerons ici que le premier cas. Les données sont stockées dans une matrice  $X$  de  $n$  lignes (observations  $x_1, \dots, x_n$ ) et de  $p$  colonnes (variables

Techniques d'enquête, juin 2008

matrice de covariance est définie positive, les données sont transformées dans l'espace des vecteurs propres de la matrice initiale. Les médianes des coordonnées et les écarts absolus par rapport à la médiane dans ce nouvel espace sont alors rétransformés dans l'espace original pour obtenir une estimation du centre et une matrice de covariance définie positive. Une autre idée de Gnanaadesikan et Kettenring (1972) est reliée aux méthodes dites de recherche avant (*forward search*), qui sont étroitement apparentées à la méthode proposée ici. Ces méthodes sont fondées sur le concept consistant à « faire pousser » un bon sous-ensemble d'observations. Par « bon sous-ensemble », il est entendu un sous-ensemble exempt ou presque exempt d'observations aberrantes. L'idée est de partir d'un petit sous-ensemble de données, puis d'ajouter des observations non aberrantes jusqu'à ce qu'il n'y en ait plus aucune disponible. La notion d'un algorithme de recherche avant a été proposée pour la première fois dans Wilks et Gnanaadesikan (1964) et décrite en détail dans Gnanaadesikan et Kettenring (1972). Les articles de Hadi (1992) et d'Atkinson (1993) ont démontré l'efficacité de ce genre de méthodes. Dans les deux articles, le « bon sous-ensemble » est agrandi un point à la fois en utilisant les distances de Mahalanobis pour classer les observations. Par la suite, les travaux de recherche ont porté surtout sur l'élaboration de méthodes plus rapides et plus perfectionnées fondées sur la même idée. Les deux dernières, qui sont les plus efficaces, ont été élaborées par Billor, Hadi et Velliaman (2000) et Kosinski (1999). Ces algorithmes ont été comparés par Béguin (2002) et l'algorithme BACON (Billor et coll. 2000) s'est avéré être la méthode de recherche avant la plus robuste et la plus rapide dans le cas de données normales multivariées complètes. En particulier, le point de défaillance s'est révélé très élevé dans les applications pratiques. En outre, comparativement à d'autres méthodes de type Mahalanobis, la performance de l'algorithme BACON sur des données complètes est très bonne (Béguin et Hülliger 2003). Aucune des méthodes susmentionnées n'est conçue pour traiter des données incomplètes issues de sondages, c'est-à-dire avec valeurs manquantes et poids de sondage. Le premier article dans lequel a été abordé le problème de la détection d'observations aberrantes multivariées dans des données incomplètes est celui de Little et Smith (1987). Ces auteurs proposent des distances de Mahalanobis pour détecter les observations aberrantes, et obtiennent des estimations du centre et de la dispersion par l'algorithme ER, dans lequel l'estimateur du maximum de vraisemblance de Lait et Rubin (1977) est remplacé par un M-estimateur robuste en une étape. Cependant, l'algorithme ER a pour point de départ la moyenne et la covariance non robustes



Dans le présent article, nous nous concentrons sur la détection des observations aberrantes. Selon notre schéma, une fois qu'une observation aberrante est détectée, elle peut être vérifiée et traitée manuellement ou elle peut être traitée par imputation. L'estimation robuste remplacerait à la fois la détection et l'imputation, mais elle est moins bien adaptée à la pratique des statistiques officielles. Nous ne faisons pas la

l'autre ne dépend des données inobservées. Dans un projet de travail du projet EUREDIT sur le développement et l'évaluation de nouvelles méthodes de vérification et d'imputation (EUREDIT 2003), les auteurs ont développé une méthode de détection des observations aberrantes pour traiter la situation difficile que représentent les données de sondage incomplètes multivariées. Deux de ces méthodes, appelées Transformed Rank Correlations et Epidemic Algorithm, sont présentées dans Béguin et Huisiger (2004). La troisième méthode, BACON-EEM, est présentée ici.

difficile de détecter une valeur aberrante qui ne l'est que dans ses valeurs manquantes. Cette situation est analogue au concept de valeur manquant au hasard (MAR, pour *missingness at random*) (Little et Rubin 1987). Nous avons besoin d'information dans les valeurs observées pour inférer qu'une observation est aberrante. Nous pourrions appeler cette situation « aberrance au hasard ». Nous pouvons l'exprimer formellement en disant que le mécanisme de création d'observations aberrantes ne dépend pas des données non observées, qui comprennent les valeurs réelles non observées de l'observation aberrante au cas où l'observation aberrante serait une erreur. Cependant, pour la détection des observations aberrantes, cette condition est trop stricte, parce que nous pourrions arriver à détecter des observations aberrantes dans des valeurs observées, même si le mécanisme dépend des valeurs non observées. Cela est possible, parce que le modèle doit tenir pour la majorité des données seulement et non pour les observations aberrantes. Si les valeurs observées d'une observation aberrante s'écartent suffisamment du modèle, cette dernière est détectée. Cependant, en ce qui concerne l'imputation des valeurs réelles pour les observations désignées aberrantes, nous sommes dans la même situation que pour les valeurs manquantes. Si le mécanisme de création des observations aberrantes, sachant les valeurs observées, dépend encore des valeurs réelles non observées de l'observation aberrante, nous ne pouvons pas estimer un modèle pour les valeurs non observées. Dans le présent article, nous utilisons l'imputation uniquement comme un outil ponctuel pour améliorer la détection des observations aberrantes. Néanmoins, nous supposons que, sachant les données observées, le mécanisme de non-réponse et le mécanisme de création d'observations aberrantes sont indépendants et que ni l'un ni

De nombreux estimateurs robustes de la matrice de covariance ont été publiés dans la littérature. Les estimations M (Huber 1981; Maronna 1976) ont l'avantage d'être relativement faciles à calculer par une itération simple à partir d'un bon point de départ (Rocke et Woodruff 1993). Toutefois, leur point de défaillance, c'est-à-dire la plus petite fraction des données dont la modification arbitraire peut mener un estimateur au-delà de toutes les bornes, est, au mieux  $1/(p+1)$ , où  $p$  est la dimension des données (Donoho 1982; Maronna 1976; Stahel 1981). Ce handicap est important dans le cas des statistiques officielles, qui sont souvent de dimensionnalité élevée. De nombreux autres estimateurs robustes équivariants affines, c'est-à-dire des estimateurs qui se transforment de manière cohérente quand les données sont transformées linéairement, ont été étudiés par Donoho (1982), mais tous ont un point de défaillance correspondant, au mieux, à  $1/(p+1)$ . D'autres approches ont abouti à des estimateurs équivariants affines à point de défaillance élevé, par exemple l'estimateur de Stahel-Donoho (SD) (Stahel 1981; Donoho 1982) ou les estimateurs par le déterminant minimal de covariance (MCD, pour *Minimum Covariance Determinant*) (Rousseeuw 1985; Rousseeuw et Leroy 1987), mais elles ont l'inconvénient d'être informatiquement coûteuses. Une approche de Gnanadesikan et Kettenring s'appuyant sur une construction de la matrice de covariance sacrifie l'équivariance affine, mais gagne en simplicité et en vitesse. Cette approche a été réactualisée dans Maronna et Zammit (2002) et dans l'une des méthodes présentées dans Béguin et Hurliger (2004), appelée corrélations de rang transformées (TRC pour *Transformed Rank Correlations*). La méthode TRC consiste à calculer une matrice initiale de corrélations de rang de Spearman bivariées. Pour s'assurer que la

grandes familles. Les auteurs de nombreuses méthodes multivariées existantes peuvent être classées en deux sous-groupes. Les uns utilisent une distance de Mahalanobis et essaient d'estimer robustement le centre et la matrice de covariance. Puis, ils utilisent une mesure de la deuxième classe de méthodes ne s'appuyant sur une hypothèse distributionnelle, mais utilisant une mesure de profondeur des données (voir Liu, Parelius et Singh 1999 pour une revue) qui sert de mesure d'aberrance. La deuxième famille est, à première vue, plus séduisante, mais malheureusement, elle produit souvent des méthodes dont les calculs ne sont pas faisables avec de grands ensembles

distinction entre les cas aberrants représentatifs et non représentatifs (Chambers 1986), puisque les uns et les autres doivent être décelés, bien qu'ils doivent éventuellement être traités différemment.



# L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées dans des données d'enquête incomplètes

Cédric Béguin et Beat Hultiger<sup>1</sup>

## Résumé

Dans le cas de données multivariées complètes, l'algorithme BACON (Billor, Hadi et Veilemann 2000) donne une estimation robuste de la matrice de covariance. La distance de Mahalanobis correspondante peut être utilisée pour la détection des observations aberrantes multivariées. Quand des items manquent, l'algorithme EEM est un moyen commode d'estimer la matrice de covariance à chaque étape d'itération de l'algorithme BACON. Dans l'échantillonnage en population finie, l'algorithme EEM doit être amélioré pour estimer la matrice de covariance de la population plutôt que de l'échantillon. Une version de l'algorithme EEM pour données d'enquête suivant un modèle normal multivarié, appelée l'algorithme EEM (espérance estimation/maximisation), est proposée. La combinaison des deux algorithmes, dénommée l'algorithme BACON-EEM, est appliquée à deux ensembles de données et comparée à d'autres méthodes.

Mots clés : Méthode de recherche avant (forward search); détection des observations aberrantes; données multivariées; valeurs manquantes; échantillonnage; robustesse; algorithme EM.

## 1. Introduction

Le problème qui sous-tend les méthodes présentées ici est celui d'une enquête par sondage portant sur des données quantitatives, comme les ventes de divers produits, pour lesquelles des valeurs manquent et des observations aberrantes existent. Souvent, à la phase de vérification du sondage, les observations aberrantes sont détectées par inspection des observations aberrantes individuelles ou par des méthodes de détection des observations aberrantes univariées. Toutefois, il existe peu de méthodes systématiques permettant de détecter des observations aberrantes multivariées dans des données d'enquête incomplètes. La détection des observations aberrantes est un aspect important de la vérification des données statistiques. Celles qui ne sont pas détectées peuvent avoir un effet important et indésirable sur les résultats de l'enquête. La plupart des méthodes existantes de détection des observations aberrantes sont conçues pour être appliquées à des données complètes univariées ou bivariées. Toutefois, les observations aberrantes réelles dans les données d'enquête sont souvent de nature multivariée. Le problème qu'elles posent devient nettement plus difficile à résoudre en trois dimensions ou plus qu'en une ou deux dimensions. Dans le contexte unidimensionnel, les observations aberrantes ne peuvent être qu'extrêmement petites ou extrêmement grandes (du moins pour les distributions unimodales), mais, dans le cas d'une dimensionnalité plus élevée, la question de la « direction » de l'observation aberrante devient de plus en plus importante. Les observations aberrantes peuvent être relativement proches de la majorité des données ou d'un modèle si la distance est mesurée dans une métrique

euclydienne, parce que cette dernière ne vérifie que les directions des axes. Par contre, si l'on utilise une métrique appropriée à la structure de corrélation de la majorité des données, l'observation aberrante peut être éloignée. Donc, pour les dimensionnalités plus élevées, la forme du nuage de points de la majorité des données doit être bien reflétée par la métrique utilisée pour pouvoir détecter les observations aberrantes.

La détection des observations aberrantes nécessite un modèle pour la majorité des données afin de pouvoir distinguer les observations auxnelles le modèle n'est pas bien ajusté. Donc, la détection des observations aberrantes est intrinsèquement liée aux modèles et à leur estimation robuste. Dans un contexte d'échantillonnage, le modèle devrait être approprié à la majorité de la population plutôt que seulement à la majorité de l'échantillon. Par conséquent, le plan d'échantillonnage devrait être pris en compte pour la détection des observations aberrantes dans des données d'enquête par sondage. Nous revenons à la discussion du rôle des poids d'échantillonnage aux sections 1 et 5.

Les données de sondage contiennent souvent des observations manquantes. La détection des observations aberrantes quand les réponses à certains items manquent doit comporter l'estimation du modèle pour la majorité des données en tenant compte de ces données manquantes. L'estimation sous valeurs manquantes sera fondée sur la relation entre les variables observées et les variables manquantes. Cette relation doit être modélisée robustement afin de la protéger contre les observations aberrantes. Si une observation est classée comme étant aberrante en se basant sur l'information complète, mais que les valeurs causant l'aberrance manquent, l'observation aberrante n'apparaîtra

1. Cédric Béguin, Université de Neuchâtel, 2010 Neuchâtel, Suisse; Beat Hultiger, Université des sciences appliquées de la Suisse du Nord-Ouest, 4600 Olten, Suisse; Courriel : beat.hultiger@fhnw.ch.

Sämdal, C.-E., Swensson, B. et Wretman, J. (1989). The weighted regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Sämdal, C.-E., Swensson, B. et Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer-Verlag.

Schlesselman, J. (1971). Power families: A note on the Box and Cox transformation. *Journal of the Royal Statistical Society, Series B*, 33, 307-311.

Skinner, C.J., Holt, D. et Smith, T.M.F. (Éds.) (1989). *Analysis of complex surveys*. Chichester : Wiley.

Spitzer, J.J. (1976). The demand for money, the liquidity trap and functional forms. *The International Economics Review*, 17, 220-227.

Sverchkov, M., et Pfeffermann, D. (2004). Prédiction des totaux de population finie basée sur la distribution échantillonnale. *Techniques d'enquête*, 30, 87-101.

Zheng, H., et Little, R.J.A. (2004). Modèles non paramétriques mixtes à fonction spline pénalisée pour l'inférence au sujet d'une moyenne de population finie d'après des échantillons à deux degrés. *Techniques d'enquête*, 30, 233-243.

Taylor, M.J.G. (1986). The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81, 114-118.

Tukey, J.W. (1957). The comparative anatomy of transformation. *Annals of Mathematical Statistics*, 28, 601-632.

Wu, C. (1999). The effective use of complete auxiliary information from survey data. Ph.D. dissertation, Simon Fraser University, Canada.

Wu, C., et Sitter, R.R. (2001). A model-calibration to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Zarembka, P. (1968). Functional form in the demand for money. *Journal of the American Statistical Association*, 63, 502-511.

- Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex survey. *Revue Internationale de Statistique*, 51, 279-92.
- Bickel, P.J., et Doksum, K.A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296-311.
- Box, G.E., et Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211-252.
- Boylan, T.A., Cuddy, M.P., et O'Muircheartaigh, I.G. (1980). The functional form of the aggregate import demand equation: A comparison of three european economies. *Journal of international economics*, 10, 561-566.
- Boylan, T.A., Cuddy, M.P., et O'Muircheartaigh, I.G. (1980). The demand equations: An application of a generalized Box-Cox methodology. *Revue Internationale de Statistique*, 50, 103-112.
- Breidt, F.J., et Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92, 831-846.
- Breidt, F.J., et Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 1026-1053.
- Carroll, R.J., et Ruppert, D. (1988). *Transformations and weighting in Regression*. Londres : Chapman and Hall.
- Cassel, C.M., Särndal, C.-E., et Wretenman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.
- Chambers, R.L., et Dorfman, A.H. (2003). Transformed variables in survey sampling. Rapport technique.
- Chen, G., et Chen, J. (1996). Une méthode de transformation applicable à l'échantillonnage de populations finies calée par une méthode de vraisemblance empirique. *Techniques d'enquête*, 22, 139-147.
- Davison, C.W., Armade, C.A., et Hallahan, C.B. (1989). Box-Cox estimation of U.S. soybean exports. *Journal of Agricultural Economics Research*, 41, 8-15.
- Deville, J.-C., et Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Estevao, V., Hidiroglou, M., et Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.
- Fuller, W.A., Loughin, M., et Baker, H. (1994). Production de poids de régression en situation de non-réponse et application à la Nationwide Food Consumption Survey de 1987-1988. *Techniques d'enquête*, 20, 79-89.
- Gemmill, G. (1980). Using the Box-Cox Form for Estimating Demand: A Comment. *The review of economics and statistics*, 62, 147-148.
- Gurka, M. (2006). Expanding the Box-Cox transformation to the linear mixed model. *Journal of the Royal Statistical Society, Series A*, 169, 273-288.
- Horvitz, D.G., et Thompson, D.J. (1952). A generalization of sampling without replacement from a finite population. *Journal of the American Statistical Association*, 47, 663-685.
- Jayasuriya, B., et Villiant, R. (1996). Application de l'estimation par régression restreinte dans une enquête-ménage. *Techniques d'enquête*, 22, 127-138.
- Jiang, J., et Lahiri, P. (2006). Estimation of finite population domain means - a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 101, 301-311.
- John, N.R., et Draper, J.A. (1980). An alternative family of transformations. *Applied Statistics*, 29, 190-197.
- Karberg, F. (2000). Survey estimation for highly skewed population in the presence of zeroes. *Journal of Official Statistics*, 16, 229-241.
- Khan, M.S., et Ross, K.Z. (1977). The functional form of aggregate import demand. *Journal of International Economics*, 7, 149-160.
- Korn, E., et Graubard, B. (1998). Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimés au moyen des données d'enquête. *Techniques d'enquête*, 24, 209-218.
- Li, Y., et Lahiri, P. (2007). Robust model-based and model-assisted predictors of the finite population total. *Journal of the American Statistical Association*, 102, 664-673.
- Miner, A.G. (1982). The contribution of weather and technology to U.S. soybean yield. *Unpublished Dissertation*. University of Minnesota.
- Montanari, G.E., et Ranalli, M.G. (2003). Nonparametric model calibration estimation in survey sampling. Manuscript.
- Montanari, G.E., et Ranalli, M.G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429-1442.
- Newman, P. (1977). Malaria and Mortality. *Journal of the American Statistical Association*, 72, 257-263.
- Royall, R.M., et Cumberland, W.G. (1981). The finite population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, 76, 924-930.
- Sakia, R.M. (1992). The Box-Cox transformation technique: A review. *The Statistician*, 41, 169-178.
- Särndal, C.-E. (1980). On  $\pi$ -inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.
- Särndal, C.-E. (1982). Implications of survey design for generalized regression estimation of linear functions. *Journal of Statistical Plan Inference*, 7, 155-170.
- Statistique Canada, N° 12-001-X au catalogue



## Remerciements

En plus, sous l'hypothèse (4), une approximation par série de Taylor de deuxième ordre de  $g_i(\theta_w)$  peut être développée comme il suit :

$$g_i(\theta_w) = g_i(\theta_N) + (\partial g_i(\mathbf{t}) / \partial \mathbf{t})' |_{\mathbf{t}=\theta_N} (\theta_w - \theta_N) + (\partial^2 g_i(\mathbf{t}) / (\partial t \partial \mathbf{t}') |_{\mathbf{t}=\theta_w} (\theta_w - \theta_N),$$

où  $\theta^* \in (\theta_N, \theta_w)$  ou  $(\theta_w, \theta_N)$ . Il découle des hypothèses (1) et (4) que

$$N^{-1} \sum_{i \in I} g_i(\theta_w) = N^{-1} \sum_{i \in I} g_i(\theta_N) + N^{-1} \sum_{i \in I} (\partial g_i(\mathbf{t}) / \partial \mathbf{t})' |_{\mathbf{t}=\theta_N} (\theta_w - \theta_N) + O^p(n^{-1}),$$

$$N^{-1} \sum_{i \in I} \pi_i^t g_i(\theta_w) = N^{-1} \sum_{i \in I} \pi_i^t g_i(\theta_N) + N^{-1} \sum_{i \in I} \pi_i^t (\partial g_i(\mathbf{t}) / \partial \mathbf{t})' |_{\mathbf{t}=\theta_N} (\theta_w - \theta_N) + O^p(n^{-1}),$$

En vertu des hypothèses (1) et (3),  $\theta_w = \theta_N + O^p(l/\sqrt{n})$  et

$$N^{-1} \sum_{i \in I} \pi_i^t \partial g_i(\mathbf{t}) / \partial \mathbf{t} |_{\mathbf{t}=\theta_N} = N^{-1} \sum_{i \in I} \partial g_i(\mathbf{t}) / \partial \mathbf{t} |_{\mathbf{t}=\theta_w}$$

$$+ O^p(l/\sqrt{n}).$$

Donc,

$$N^{-1} \left( \sum_{i \in I} \pi_i^t g_i(\theta_w) - \sum_{i \in I} \pi_i^t g_i(\theta_N) \right) = N^{-1} \left( \sum_{i \in I} g_i(\theta_N) - \sum_{i \in I} \pi_i^t g_i(\theta_N) \right) + O^p(n^{-1}).$$

Par conséquent,

$$\hat{T}_{G-BC}^p = \sum_{i \in S} Y_i / \pi_i$$

$$+ \left( \sum_{i \in I} g_i(\theta_N) - \sum_{i \in I} \pi_i^t g_i(\theta_N) \right) + O^p(N/n).$$

La variance asymptotique de  $\hat{T}_{AG}$  sous le plan est :

$$AV^p(\hat{T}_{G-BC}^p) \approx \sum_{i \in I} \sum_{j \in I} (\pi_i^t - \pi_i \pi_j) (Y_i - g_i(\theta_N)) (Y_j - g_j(\theta_N)) / (\pi_i \pi_j),$$

qui peut être estimée par

$$\hat{V}_{G-BC}^p$$

$$\approx \sum_{i \in S} \sum_{j \in S} (\pi_i^t - \pi_i \pi_j) (Y_i - g_i(\theta_N)) (Y_j - g_j(\theta_N)) / (\pi_i \pi_j),$$

## Annexe

Soit

$$\hat{T}_{G-BC}^p = \sum_{i \in S} Y_i / \pi_i + \left( \sum_{i \in I} g_i(\theta_w) - \sum_{i \in I} \pi_i^t g_i(\theta_w) / \pi_i^t \right).$$

Par développement en série de Taylor, sous l'hypothèse (2),

nous avons

$$g_i(\theta_w) = g_i(\theta_N) + (\partial g_i(\mathbf{t}) / \partial \mathbf{t})' |_{\mathbf{t}=\theta_w} (\theta_w - \theta_N),$$

où  $\theta^* \in (\theta_N, \theta_w)$  ou  $(\theta_w, \theta_N)$  et  $(\partial g_i(\mathbf{t}) / \partial \mathbf{t})'$  est un vecteur ligne.

En vertu des hypothèses (1) et (2),

$$N^{-1} \sum_{i \in I} g_i(\theta_w) = N^{-1} \sum_{i \in I} g_i(\theta_N) + O^p(l/\sqrt{n})$$

$$N^{-1} \sum_{i \in I} \pi_i^t g_i(\theta_w) = N^{-1} \sum_{i \in I} \pi_i^t g_i(\theta_N) + O^p(l/\sqrt{n}).$$

Notons aussi qu'en vertu de l'hypothèse (3),

$$N^{-1} \sum_{i \in I} g_i(\theta_N) = N^{-1} \sum_{i \in I} \pi_i^t g_i(\theta_N) + O^p(l/\sqrt{n}).$$

Donc,

$$N^{-1} \left( \sum_{i \in I} g_i(\theta_w) - \sum_{i \in I} \pi_i^t g_i(\theta_w) \right) = O^p(l/\sqrt{n}).$$

En outre, en vertu de l'hypothèse (3),

$$N^{-1} \left( \sum_{i \in I} Y_i - \sum_{i \in I} \pi_i^t Y_i \right) = O^p(l/\sqrt{n}).$$

Par conséquent,  $N^{-1}(\hat{T}_{G-BC}^p - T) = O^p(l/\sqrt{n})$ , i.e.,  $\hat{T}_{G-BC}^p$  converge en probabilité vers  $T$  à l'ordre  $O^p(l/\sqrt{n})$ .

**Tableau 4**  
Biais relatif et racine de l'erreur quadratique moyenne de  $\hat{\lambda}_1$  et  $\hat{\lambda}_w^2$  sous échantillonnage EAS pour diverses tailles d'échantillon et divers écarts-types

	$\sigma = 2$			$\sigma = 1$			$\sigma = 0,5$
	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}_w^2$	$\hat{\lambda}$	$\hat{\lambda}_w$	$\hat{\lambda}_w^2$	$\hat{\lambda}_w$
Biais relatif							
$n = 30$	-0,58	0,13	-0,28	0,10	-0,16	-0,01	
$n = 80$	-0,42	0,14	-0,19	0,11	-0,13	-0,02	
$n = 150$	-0,39	0,10	-0,16	0,09	-0,10	-0,01	
Racine de l'erreur quadratique moyenne							
$n = 30$	0,14	0,12	0,11	0,07	0,07		
$n = 80$	0,08	0,07	0,06	0,04	0,04		
$n = 150$	0,06	0,05	0,05	0,04	0,03	0,03	
<sup>1</sup> L'estimateur $\hat{\lambda}$ est obtenu par la méthode des moindres carrés ordinaires/méthode du maximum de vraisemblance.							
<sup>2</sup> L'estimateur $\hat{\lambda}_w^2$ est obtenu par la méthode du pseudo maximum de vraisemblance.							

## 5. Conclusion

Nous avons proposé un estimateur par la régression généralisée d'un total de population finie fondé sur une technique de transformation de Box-Cox sous un plan d'échantillonnage à probabilités inégales. Étant convergent par rapport au plan, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est détaillant. Dans de nombreuses situations, une version du modèle  $M_3$  fournira au moins une approximation utile du comportement de la variable dépendante. La technique de Box-Cox permet qu'une transformation raisonnablement ajustée de la variable dépendante soit automatiquement déterminée par les données. Nous avons évalué analytiquement et par simulation de Monte Carlo la robustesse et l'efficacité de l'estimateur proposé. Lorsqu'on compare un estimateur GREG basé sur un modèle linéaire sous-jacent ( $T_{G-L}$ ) à un autre fondé sur un modèle de Box-Cox ( $T_{G-BC}$ ), il ne faut pas perdre de vue que  $T_{G-L}$  ne requiert pas d'information auxiliaire complète. En outre, il peut produire un jeu unique de poids applicable à toutes les variables d'intérêt, contrairement à l'estimateur  $T_{G-BC}$ . Toutefois, pour accroître l'efficacité de l'un et l'autre estimateur, il faut habituellement utiliser des poids distincts pour des variables d'intérêt différentes, parce les meilleures prédictions sont obtenues en ajustant un modèle de travail particulier pour chaque variable étudiée. L'estimateur  $T_{G-BC}$  peut même être plus efficace que l'estimateur  $T_{G-L}$  mais au prix de nécessiter des données complètes sur les variables  $x$ . Bien que ce genre d'information soit rarement disponible dans les sondages nord-américains auprès des ménages, elle l'est souvent dans les sondages auprès des entreprises.

$$\mathbf{p} = \left( \sum_{i \in S} \mathbf{w}_i \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \left( \sum_{i \in S} \mathbf{w}_i \mathbf{z}_i y_i \right), \text{ et } \mathbf{z}_i = (1 \ y_i \ w_i)'$$
$$(w_i^c)' = [1 + \left( \sum_{i \in U} \mathbf{z}_i' / \pi_i \right) \left( \sum_{i \in S} \mathbf{z}_i' / \pi_i \right)^{-1} \mathbf{z}_i' ] w_i'$$

produira un ensemble de poids de calage

$$\hat{T}_{G-BC} = \sum_{i \in S} y_i / \pi_i + \left( \sum_{i \in U} \mathbf{z}_i' / \pi_i - \sum_{i \in S} \mathbf{z}_i' / \pi_i \right) \mathbf{p},$$

rapport au plan, de sorte que

$\hat{y}_{i,w}$  tirée de l'équation (3) comme la variable auxiliaire modèle (Wu et Sitter 2001). Le fait de traiter la prédiction autre approche consiste à utiliser le calage fondé sur un d'échantillonnage standard décrit dans Li et Lahiri (2007). Une estimations peuvent être obtenues en utilisant un outil estimateurs de Box-Cox ne sont pas linéaires en  $y$ . Ces accorder une attention particulière à ce problème, car les total fini pour l'ensemble de la population. Nous devons somme de ces estimations concorde avec l'estimation du lement exclusives et exhaustives de telle façon que la estimer les totaux finis pour des sous-population mutuel-seule variable d'intérêt. La question est de savoir comment Les sondages sont peu souvent réalisés pour mesurer une

Dans le présent article, nous n'avons transformé que la variable  $y$  par la technique de Box-Cox. Dans l'avenir, notre méthode pourrait être étendue à diverses formes fonctionnelles de la transformation de Box-Cox.

L'estimateur de variance de l'estimateur  $T_{G-BC}$  proposé ici est fondé sur une approximation en grand échantillon. Certaines études ont révélé des propriétés médiocres pour ce type d'estimateur de variance sous certains choix des modèles auxiliaires pour l'estimateur  $T_{G-L}$ . Les études théoriques et empiriques de la propriété de couverture de l'estimateur de variance proposé doivent se poursuivre.

LI : Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox

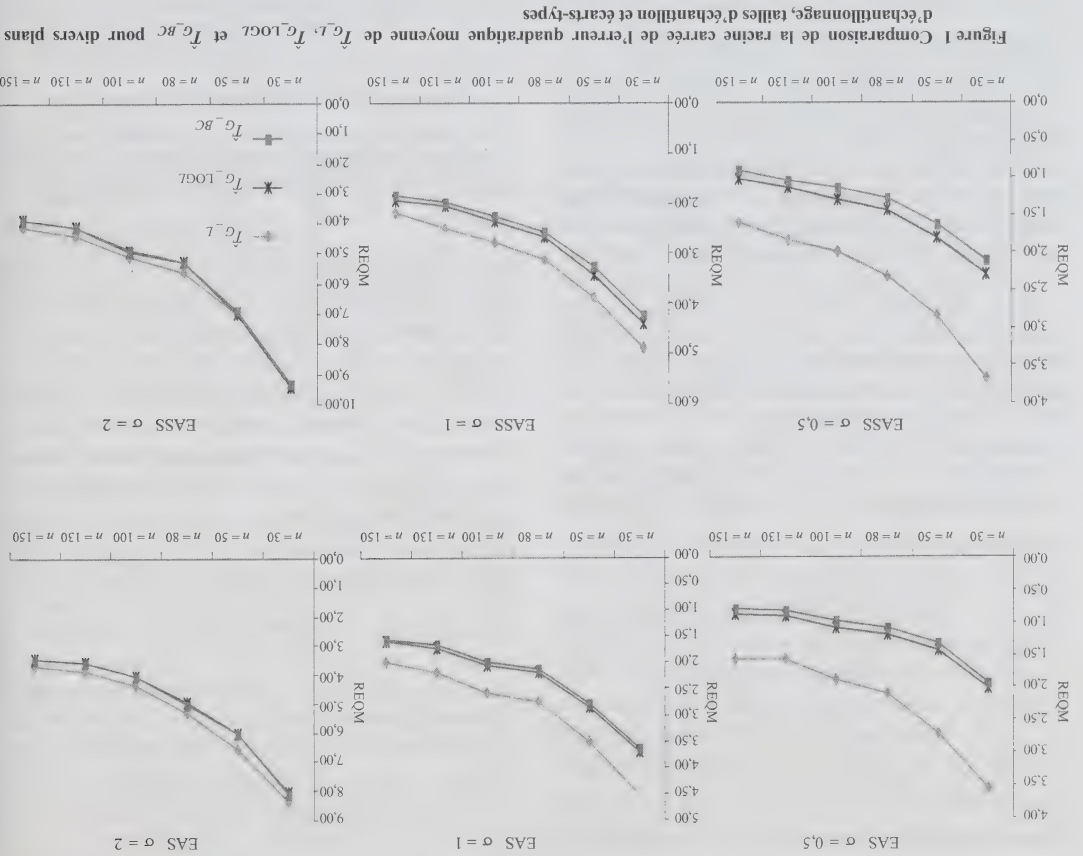


Tableau 2

Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon ( $\lambda = 1$ )

$T_1^D$	$T_2^L$	$T_3^L$	$T_4^{BC}$
Echantillonage aléatoire simple	BiaisRel( $\times 10^{-7}$ )	REQM( $\times 10^7$ )	Echantillonage aléatoire simple
$n = 30$	0,31	20,91	0,16
$n = 80$	-0,37	-0,09	0,03
$n = 150$	-0,25	0,03	0,10
$n = 30$	0,31	20,91	0,16
$n = 80$	-0,37	-0,09	0,03
$n = 150$	-0,25	0,03	0,10
REQM( $\times 10^7$ )	$n = 30$	20,91	0,16
$n = 80$	-0,37	-0,09	0,03
$n = 150$	-0,25	0,03	0,10
BiaisRel( $\times 10^{-7}$ )	$n = 30$	0,31	0,16
$n = 80$	-0,37	-0,09	0,03
$n = 150$	-0,25	0,03	0,10
Echantillonage aléatoire simple	BiaisRel( $\times 10^{-7}$ )	REQM( $\times 10^7$ )	Echantillonage aléatoire simple
$n = 30$	0,51	0,27	0,56
$n = 80$	3,97	-0,11	-0,13
$n = 150$	-0,23	0,04	0,06
$n = 30$	12,54	3,79	4,20
$n = 80$	8,39	2,27	2,61
$n = 150$	5,48	1,67	1,90
Echantillonage aléatoire simple	BiaisRel( $\times 10^{-7}$ )	REQM( $\times 10^7$ )	Echantillonage aléatoire simple
$n = 30$	0,51	0,27	0,56
$n = 80$	3,97	-0,11	-0,13
$n = 150$	-0,23	0,04	0,06
$n = 30$	12,54	3,79	4,20
$n = 80$	8,39	2,27	2,61
$n = 150$	5,48	1,67	1,90
Echantillonage aléatoire simple	BiaisRel( $\times 10^{-7}$ )	REQM( $\times 10^7$ )	Echantillonage aléatoire simple
$n = 30$	0,51	0,27	0,56
$n = 80$	3,97	-0,11	-0,13
$n = 150$	-0,23	0,04	0,06
$n = 30$	12,54	3,79	4,20
$n = 80$	8,39	2,27	2,61
$n = 150$	5,48	1,67	1,90

Tableau 3

Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon (valeurs de  $y$  produites à partir d'un modèle  $M_0$ )



Nous notons que  $\hat{T}_{GBC}$  possède systématiquement la plus petite REQM quand  $\sigma = 0,5$  et 1. Donc, un modèle robuste choisi par la méthode de Box-Cox réduit la REQM, surtout quand le modèle est approprié pour un petit  $\sigma$ .

Nous pourrions aussi nous intéresser aux propriétés de  $\hat{T}_{GL}$  comparativement à  $\hat{T}_{GBC}$  quand  $\lambda = 1$ , c'est-à-dire la situation qui favorise  $\hat{T}_{GL}$ . Si nous examinons le tableau 2, nous voyons que  $\hat{T}_{GBC}$  est relativement comparable à  $\hat{T}_{GL}$  surtout quand la taille d'échantillon est grande. Ce résultat implique que le fait d'utiliser  $\hat{T}_{GBC}$  alors que nous devrions utiliser  $\hat{T}_{GL}$  n'entraîne pas une grande perte en ce qui concerne le biais relatif et la REQM.

Afin de mieux évaluer la robustesse et l'accroissement de l'efficacité de  $\hat{T}_{GBC}$  comparativement à  $\hat{T}_{GL}$  et à  $\hat{T}_{GLOGL}$ , nous produisons une population finie de taille  $N = 4\,000$ , non plus conformément au modèle de Box-Cox  $M_5$ , mais à partir de

$$M_6: y_i^{(v)} = \beta_0 + \beta_1 x_i + z_i + e_i$$

où  $z_i = x_i^2$ . Nous utilisons les mêmes  $x_i$  et les mêmes valeurs de paramètres que celles spécifiées à la section 3 pour générer les  $y_i$ . Nous étudions les quatre mêmes estimateurs basés sur la nouvelle population finie  $\{(x_i, y_i) : i = 1, \dots, 4\,000\}$ . Il s'agit de la situation non idéale pour n'importe lequel des estimateurs GREG. Les résultats sont présentés au tableau 3. Les avantages de l'utilisation de  $\hat{T}_{GBC}$  sont manifestes en ce qui concerne le biais relatif et la REQM.

Le tableau 4 donne le biais relatif et la REQM pour  $\hat{\lambda}$  et  $\hat{\lambda}_w$  sous échantillonnage EASS avec diverses tailles d'échantillon et les  $\sigma$  fondés sur les valeurs de population produites à partir de  $M_5$ . Puisque nous avons stratifié sur la variable  $y$ , l'utilisation des pondérations devrait avoir un impact, du moins sur le biais de l'estimation des paramètres. Quand  $\sigma$  est petit, c'est-à-dire quand les données simulées sont bien ajustées au modèle hypothétique,  $\hat{\lambda}_w$  donne un biais relatif plus proche de 0, mais  $\hat{\lambda}_w$  et  $\hat{\lambda}$  donnent l'un et l'autre d'aussi bons résultats en ce qui concerne la REQM. Par contre, quand  $\sigma$  est grand,  $\hat{\lambda}_w$  donne systématiquement des valeurs absolues plus petites du biais relatif et de la REQM que  $\hat{\lambda}$ , quoique les REQM restent proches. En effet, en ce qui concerne l'estimation de  $T$ , ni l'une ni l'autre approche ne l'importe pour ce qui est du biais empirique ou de la racine carrée de l'erreur quadratique moyenne, quelle que soit la taille de l'échantillon ou la valeur choisie de  $\sigma$  (données non présentées). Cela pourrait tenir au fait que l'estimateur de  $T$  est biaisé sous le modèle quelle que soit la qualité de l'estimation de  $\lambda$ .

où  $B$  est le nombre de répétitions dans la simulation de Monte Carlo et  $\hat{w}$  représente une estimation arbitraire du paramètre de population finie  $w$ .

$$REQM = \sqrt{B \sum_{b=1}^B (\hat{w}^b - w)^2}$$

#### 4. Résultats

Au tableau 1, nous présentons le biais relatif et la racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillon quand  $\sigma = 0,5$ . Les quatre estimateurs donnent tous un biais relatif proche de 0 [le maximum des valeurs absolues du biais relatif (BiaisRel)] est inférieur à 0,01 dans le tableau 1]. Parmi ces estimateurs,  $\hat{T}_{GBC}$  est celui dont le biais relatif absolu et la REQM sont les plus faibles sur divers tailles d'échantillon et plans d'échantillonnage. Par conséquent, la technique de Box-Cox protège  $\hat{T}_{GBC}$  qui voit son efficacité accrue comparativement à d'autres estimateurs GREG.

**Tableau 1**  
Biais relatif et racine carrée de l'erreur quadratique moyenne de quatre estimateurs pour différents plans d'échantillonnage et diverses tailles d'échantillon ( $\lambda = 0,1$ )

Echantillonnage aléatoire simple

BiaisRel( $\times 10^{-3}$ )

REQM( $\times 10^3$ )

$n = 30$

$n = 80$

$n = 150$

$\hat{T}_D$

$\hat{T}_{GL}$

$\hat{T}_{GLOGL}$

$\hat{T}_{GBC}$

BiaisRel( $\times 10^{-3}$ )

REQM( $\times 10^3$ )

$n = 30$

$n = 80$

$n = 150$

Echantillonnage aléatoire simple stratifié

BiaisRel( $\times 10^{-3}$ )

REQM( $\times 10^3$ )

$n = 30$

$n = 80$

$n = 150$

estimateur d'Horvitz-Thompson fondé sur le plan;

estimateur GREG avec modèle linéaire sous-jacent;

estimateur GREG avec modèle logarithmique sous-jacent;

estimateur GREG avec modèle de Box-Cox sous-jacent.

Nous étudions aussi le biais relatif et la REQM sous les mêmes conditions quand  $\sigma = 1$  et  $\sigma = 2$ . La figure 1 donne la REQM pour les trois estimateurs GREG ( $\hat{T}_{GL}$ ,  $\hat{T}_{GLOGL}$  et  $\hat{T}_{GBC}$ ) lorsqu'on utilise différents plans d'échantillonnage.

moyenne  $\mu_x$  et d'erreur-type  $\sigma_x$ . Afin d'imiter une situation réelle, nous choisissons  $\lambda = 0,1$ ,  $\beta_0 = 4,20$  et  $\beta_1 = 2,66$  qui sont les estimations obtenues en ajustant les données réelles d'enquête au modèle  $M_3$ . Nous fixons  $\mu_x = 1\ 040$  et  $\sigma_x = 1\ 000$  pour être certain que  $y_i > 0$  pour presque toutes les unités  $i$ . Strictement parlant, nous avons une distribution normale tronquée de  $y_i$ , puisque nous écartons toutes les valeurs négatives. L'effet en est négligeable, car moins de 0,1 % des valeurs produites de  $y_i$  doivent être supprimées. Le même phénomène a été observé par Taylor (1986).

La simulation est fondée sur un échantillonnage répété à partir de la population finie générale. Nous étudions deux plans d'échantillonnage, à savoir l'échantillonnage aléatoire simple (EAS) et l'échantillonnage aléatoire simple stratifié (EASS). Lorsqu'un échantillon est sélectionné par EASS, nous appliquons des probabilités de sélection inégales entre les diverses strates. Nous définissons deux strates en utilisant comme valeur limite la médiane des valeurs de  $y$  dans la population finie. Pour la strate  $h$  de taille  $N_h$ , nous sélectionnons un échantillon aléatoire simple de taille  $n_h$ . Définissons les probabilités de sélection  $p_1$  et  $p_2$  pour la strate 1 et la strate 2, respectivement. Nous spécifions  $p_1 = 2 \times p_2$ . Pour une taille d'échantillon fixe  $n$ ,  $n_1 = N_1 \times p_1$  et  $n_2 = N_2 \times p_2$ . Nous souhaitons estimer le total de population finie

$$T = \sum_{i \in U} y_i.$$

Dans cette étude par simulation, nous étudions les propriétés des estimateurs  $\hat{T}_{G,BC}$ ,  $\hat{T}_{G,LP}$  et  $\hat{T}_{G,LOG}$ , ainsi que l'estimateur d'Horvitz-Thompson fondé sur le plan  $(\hat{T}_D)$ , où les indices inférieurs «-L», «-LOGL» et «-BC» dénotent les modèles linéaire, logarithmique et Box-Cox sous-jacents, respectivement.

Un millier d'échantillons sont tirés de la population finie simulée pour chaque taille d'échantillon  $n \in \{30, 80, 150\}$ . Quatre estimateurs sont produits pour chaque échantillon sélectionné. L'estimateur du paramètre de transformation en population finie  $\lambda$  est également produit pour chaque échantillon. Pour les besoins de la comparaison, nous utilisons deux méthodes pour estimer  $\lambda$ . Soit  $\hat{\lambda}(\hat{\lambda}_m)$  les estimateurs MCO/MV (PMV) de  $\lambda$ . Sur l'ensemble des 1 000 échantillons, nous calculons les biais relatifs empiriques en pourcentage (BiaisRel) et les racines carrées des erreurs quadratiques moyennes (REQM) pour évaluer ces estimateurs en utilisant les formules suivantes :

$$\text{BiaisRel} = B^{-1} \sum_{b=1}^B (\hat{\omega}_b - \omega) / \omega,$$

Hypothèse 2 : Pour chaque  $x_i$ ,  $\partial g_i(t) / \partial t \leq h(t)$  pour  $t$  dans le voisinage de  $\theta$ .

Hypothèse 3 : Les estimateurs d'Horvitz-Thompson calculés avec les poids de sondage de base pour certains totaux de population suivent une loi asymptotiquement normale.

Hypothèse 4 : Pour chaque  $x_i$ , la dérivée seconde de  $g_i(t)$  par rapport à  $t$  est continue et bornée dans le voisinage de  $\theta$ .

Preuve : (voir l'annexe).

L'estimateur proposé de la variance représenté par l'équation (5) est basé sur des approximations en grand échantillon. Pour un niveau nominal donné  $1 - \alpha$ , l'intervalle de confiance habituel fondé sur l'approximation normale pour l'estimateur de la variance donne approximativement un taux de couverture de  $100(1 - \alpha)\%$  dans les grands échantillons répétés. Malheureusement, dans certains cas, on a constaté que les propriétés de couverture de ce type d'estimateur de la variance peuvent être médiocres pour certains choix du modèle auxiliaire (*assisting model*) pour les estimateurs  $\hat{T}_{G,L}$  (Sämdal 1982; Sämdal, Swensson et Wretman 1989; etc.). Les études théoriques et empiriques de la propriété de couverture de l'estimateur proposé de la variance devront être approfondies.

L'estimateur  $\hat{T}_{G,BC}$  est convergent par rapport au plan pour le total de population finie  $T$  sous l'approche de randomisation, et la technique de Box-Cox permet qu'une transformation raisonnable de la variable dépendante soit déterminée automatiquement par les données à partir d'une grande famille de fonctions, donc permet d'acroître l'efficacité.

### 3. Une étude par simulation

L'objectif de cette étude par simulation est d'évaluer les propriétés de divers estimateurs GREG pour un total de population finie. Dans cet exercice de simulation, nous produisons une population finie à partir de l'Australian Agricultural and Grazing Industries Survey (AAGIS). Ces données d'enquête contiennent des renseignements sur le nombre de têtes de bétail ( $y$ ) et la superficie agricole ( $x$ ) pour chacune des 431 fermes.

Nous considérons une population finie de taille  $N = 4\ 000$ , générée à partir du modèle suivant :

$$M_5: y_i^{(\alpha)} = (y_i^j)^{1/\lambda} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

où les  $\varepsilon_i$  sont indépendants et de loi approximativement normale  $N(0, \sigma^2)$ , et  $x_i$  est le logarithme d'une valeur produite à partir d'une distribution exponentielle de

$$l_i(\Phi^*) = \sum_{i \in I} [\partial \log f(Y_i; \Phi^*, Y) / \partial \Phi^*] = 0,$$

où

$$\tilde{Y} = \prod_{N=1}^I Y_i^{1/N}.$$

Pour  $\Phi^*$  donné, posons que  $l_i(\Phi^*)$ , sommation de la dérivée première de la log-vraisemblance par rapport à  $\Phi^*$ , est un paramètre de population finie. Nous tirons un échantillon  $i$  en approximant  $\log f(Y_i; \Phi^*, Y)$  pour chaque unité  $i$  de l'échantillon par  $\log f(Y_i; \Phi^*, Y^w)$ , nous estimons le total de population,  $l_i^s(\Phi^*)$ , par  $l_i^s(\Phi^*)_{PMV}$  :

$$l_i^s(\Phi^*)_{PMV} = \sum_{i \in I} w_i [\partial \log f(Y_i; \Phi^*, Y^w) / \partial \Phi^*]_{\Phi^* = \Phi^*_{PMV}},$$

où

$$\tilde{Y}^w = \prod_{i \in I} Y_i^{w_i / \sum w_i},$$

la moyenne géométrique pondérée des  $Y$  dans l'échantillon et  $\Phi^*_{PMV}$  est l'estimateur du pseudo maximum de vraisemblance de  $\Phi^*$ , qui satisfait  $l_i^s(\Phi^*_{PMV}) = 0$  (Wu et Sitter 2001). L'estimateur  $PMV$ ,  $\Phi^*_{PMV} = (\hat{\beta}^w, \hat{\gamma}^w, \hat{\sigma}^w_{e_2})$ , peut être obtenu par une méthode de balayage de l'espace des solutions. Autrement dit, calculer et tracer les valeurs de log-vraisemblance pondérées,

$$(4) \quad \log L(\Phi^*) = \sum_{i \in I} w_i \log f(Y_i; \Phi^*, Y^w)$$

et

$$\hat{T}_{G-BC} = \sum_{i \in I} g_i(\theta^w) + \sum_{i \in I} (Y_i - g_i(\theta^w)) / \pi_i.$$

**Théorème :** Sous les hypothèses suivantes, l'estimateur GREG fondé sur la transformation de Box-Cox  $\hat{T}_{G-BC}$  est convergent par rapport au plan pour  $T$ , en ce sens que asymptotique de  $\hat{T}_{G-BC}$  est donnée par

$$AV_A(\hat{T}_{G-BC}) = \sum_{i \in I} \sum_{j \in I} (\pi_{ij} - \pi_i \pi_j) (Y_i - g_i(\theta^N)) (Y_j - g_j(\theta^N)) / (\pi_i \pi_j),$$

qui peut être estimée par

$$\hat{V}_A(\hat{T}_{G-BC}) = \sum_{i \in I} \sum_{j \in I} ((\pi_{ij} - \pi_i \pi_j) / \pi_{ij}) (Y_i - g_i(\theta^w)) (Y_j - g_j(\theta^w)) / (\pi_i \pi_j). \quad (5)$$

$$\text{Hypothèse 1 : } \theta^w = \theta^N + O_p(1/\sqrt{n}).$$



et aux variables  $x$ . Li et Lahiri (2007) ont également utilisé le modèle général de Box-Cox pour prédire le total de population finie sous un cadre fondé sur un modèle. Dans l'avenir, cette méthode pourrait être étendue à diverses formes fonctionnelles de la transformation de Box-Cox. Ici, nous discutons uniquement d'une transformation Box-Cox de la variable  $y$ , ce qui permet une comparaison plus équitable des trois estimateurs GREG.

### 2.2 Estimation des paramètres du modèle et de la transformation $\phi = (\beta, \lambda, \sigma^2)'$ par la méthode du pseudo maximum de vraisemblance (PMV)

Afin de faciliter l'estimation de  $\lambda$  en utilisant les procédures de calcul existantes, il faut remplacer  $\mathbf{Y}^{(A)}$  dans le modèle  $\mathbf{M}_3$  par une transformation à l'échelle  $\mathbf{Y}^{*(A)}$ . Pour la  $i^{\text{e}}$  unité,

$$y_i^{*(A)} = \begin{cases} y_i \log(y_i) & \lambda = 0, \\ (y_i^\lambda - 1) / \lambda y_i^{\lambda-1} & \lambda \neq 0, \end{cases}$$

où  $y_i$  est la moyenne géométrique des  $y_i$ . Les calculs qui suivent seront fondés sur le nouveau modèle à l'échelle :

$$\mathbf{M}_4 : \mathbf{Y}^{*(A)} \mathbf{X} \boldsymbol{\beta}^* = \boldsymbol{\varepsilon}^*$$

où les  $\boldsymbol{\varepsilon}^*$  suivent une loi approximativement normale de moyenne  $\mathbf{0}$  et de matrice de variance  $\sigma^{\varepsilon^*} \mathbf{I}$ . Soit  $\boldsymbol{\phi}^* = (\boldsymbol{\beta}^*, \lambda, \sigma^{\varepsilon^*})'$ .

L'estimateur du maximum de vraisemblance (EMV) de  $\boldsymbol{\phi}^*$  maximise la log-vraisemblance

$$l(\boldsymbol{\phi}^*) = \sum_{i=1}^I \log f(y_i; \boldsymbol{\phi}^*, y_i),$$

où

$$f(y_i; \boldsymbol{\phi}^*, y_i) = (2\pi\sigma^{\varepsilon^*})^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y}_i^{*(A)} - \mathbf{x}_i' \boldsymbol{\beta}^*)^2 / \sigma^{\varepsilon^*}\right\} \cdot (y_i / y_i)^{\lambda-1}.$$

Skinner, Holt et Smith (1989) redéfinissent  $\boldsymbol{\phi}^*$  comme la valeur de  $\boldsymbol{\phi}^*$  qui maximise

$$l(\boldsymbol{\phi}^*) = \sum_{i \in U} \log f(y_i; \boldsymbol{\phi}^*, \boldsymbol{\phi}^*),$$

la somme étant calculée sur l'ensemble des unités de la population finie. Donc, parmi tous les modèles possibles  $f(y_i; \boldsymbol{\phi}^*, \boldsymbol{\phi}^*)$ , on choisit celui qui « est le mieux ajusté » à la population finie. Si nous choisissons mal la famille  $f(y_i; \boldsymbol{\phi}^*, \boldsymbol{\phi}^*)$ , cet ajustement le meilleur restera médiocre, mais dans notre inférence, nous le traitons comme la cible que nous essayons d'atteindre avec nos données d'échantillon. Par conséquent, il est important de choisir les options appropriées pour  $f(y_i; \boldsymbol{\phi}^*, \boldsymbol{\phi}^*)$ . Pour la population finie,  $\boldsymbol{\phi}^*$  satisfait

où les résidus  $\boldsymbol{\varepsilon}$  suivent une loi approximativement normale de moyenne  $\mathbf{0}$  et de matrice de variance  $\sigma^2 \mathbf{I}$ .

Schleselman (1971) a montré que les estimateurs du maximum de vraisemblance des paramètres du modèle de Box-Cox sont indépendants de l'échelle, si bien que le même fonction de log-vraisemblance à condition que le modèle de régression contienne un terme constant. Inspirés par cette étude et par la plupart des articles sur les modèles de Box-Cox, nous incluons un terme constant dans le modèle.

Sous  $\mathbf{M}_3$ , le prédicteur pour la  $i^{\text{e}}$  unité non observée s'obtient par une simple rétrotransformation à partir de la transformation de Box-Cox :

$$y_i^{*w} = g_i(\boldsymbol{\beta}^{*w}, \lambda^{*w}) = (\lambda^{*w} \mathbf{x}_i' \boldsymbol{\beta}^{*w} + 1)^{1/\lambda^{*w}}, \quad (3)$$

où  $\boldsymbol{\beta}^{*w}$  et  $\lambda^{*w}$  sont les estimateurs des paramètres du modèle. La méthode d'estimation est expliquée à la section suivante. Les équations (2) et (3) ne fournissent pas de prédicteur sans biais de  $y_i$  pour les modèles sous-jacents respectifs. Li et Lahiri (2007) ont montré que, si la variance des résidus est faible, c'est-à-dire si le modèle est très bien ajusté aux données, les deuxièmes membres des équations (2) et (3) sont de bonnes solutions de rechange pour les prédicteurs sans biais. Pour simplifier et pour réduire le fardeau des calculs, nous traiterons les deuxièmes membres des équations (2) et (3) comme des alternatives appropriées pour les prédicteurs sans biais. Rappelons que notre but est de décrire des estimateurs assistés par modèle pour des totaux de population finie. Le modèle sous-jacent est utilisé uniquement pour suggérer un estimateur, qui sera évalué dans le cadre de randomisation. Même si les prédicteurs des individus sont biaisés, nous pouvons encore construire des estimateurs convergents par rapport au plan pour des totaux de population finie, contrairement aux estimateurs strictement fondés sur un modèle proposé par Li et Lahiri (2007). Nous dénotons les estimateurs GREG sous les trois modèles  $\mathbf{M}_1$  à  $\mathbf{M}_3$  par  $\hat{T}_{G-L}$ ,  $\hat{T}_{G-BC}$ , et  $\hat{T}_{G-LOG}$ , respectivement. L'estimateur  $\hat{T}_{G-BC}$  diffère des estimateurs  $\hat{T}_{G-L}$  et  $\hat{T}_{G-LOG}$ , parce que les données dicte la transformation qu'il faut utiliser.

Il est possible d'intégrer des transformations de Box-Cox sur la variable  $y$  ainsi que les variables  $x$ . Par le passé, diverses formes fonctionnelles du modèle de Box-Cox ont été étudiées. Khan et Ross (1977), Spitzer (1976), Zarembka (1968), Boylan, Cuddy et O'Muircheartaigh (1980), et d'autres ont considéré un cas particulier du modèle général de Box-Cox quand on émet l'hypothèse d'un paramètre de transformation commun pour la variable  $y$  et les variables  $x$ . Gemmill (1980), Boylan, Cuddy et O'Muircheartaigh (1982), et d'autres ont appliqué le modèle général de Box-Cox avec divers paramètres de transformation à la variable  $y$ .

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \mathbf{Y}_3 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{pmatrix},$$

où

- $\mathbf{Y}_1$  est un vecteur colonne de dimensions  $n \times 1$  de valeurs observées de la variable étudiée;
- $\mathbf{Y}_2$  est un vecteur colonne de dimensions  $(N - n) \times 1$  de valeurs non observées de la variable étudiée;
- $\mathbf{X}_1$  est une matrice de dimensions  $n \times (k + 1)$  de variables auxiliaires connues dans l'échantillon;
- $\mathbf{X}_2$  est une matrice de dimensions  $(N - n) \times (k + 1)$  de variables auxiliaires connues hors de l'échantillon.

Tout au long de l'article, nous utilisons  $E_d$  et  $V_d$  pour dénoter l'espérance, ou valeur prédite, et la variance sous le plan d'échantillonnage.

## 2.1 Estimateurs GREG de totaux de population finie

L'estimateur GREG est défini ici comme étant

$$\hat{T}_G = \sum_{i \in I} \hat{y}_{i,w} + \left( \sum_{i \in I} (y_i - \hat{y}_{i,w}) / \pi_i \right)$$

où  $\hat{y}_{i,w}$  est le prédicteur de  $y_i$  fondé sur un modèle. Quel

que soit le degré avec lequel le modèle sous-jacent décrit bien la population, les estimateurs GREG du total de population finie sont convergents par rapport au plan sous des contraintes faibles (Sæmdal et coll. 1992). Le modèle utilisé le plus fréquemment est le modèle de régression linéaire standard, donné par

$$\mathbf{M}_1 : \mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

où  $\varepsilon \sim (0, \sigma^2 \mathbf{I})$ , une loi de probabilité à  $N$  variables de vecteur moyen  $\mathbf{0}$  et de matrice de variance-covariance  $\sigma^2 \mathbf{I}$ , et  $\mathbf{I}$  est la matrice d'identité de dimensions  $N \times N$  (rien ne serait perdu dans ce contexte en remplaçant  $\sigma^2 \mathbf{I}$  par une matrice définie positive plus générale). Dans cette équation,  $\beta$  est un vecteur colonne de dimensions  $(k + 1) \times 1$  de coefficients de régression. Les paramètres  $\sigma^2$  et  $\beta$  sont l'un et l'autre des paramètres de superpopulation inconnus. Un prédicteur sans biais pour la  $i^{\text{e}}$  unité est donné par

$$(1) \quad \hat{y}_{i,w} = \mathbf{x}_i' \hat{\beta}$$

où  $\hat{\beta}$  est l'estimateur des moindres carrés pondérés (MCP) de  $\beta$  sous  $\mathbf{M}_1$  et

$$\hat{\beta}_w = \left( \sum_{i \in I} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i \in I} w_i \mathbf{x}_i y_i \right).$$

Dans certaines applications, surtout dans les enquêtes auprès des entreprises et les enquêtes agricoles, un modèle linéaire n'est pas nécessairement approprié pour  $y_i$ , mais peut être raisonnable pour une transformation strictement monotone de  $y_i$ . Pour l'ensemble de données utilisé dans

$$(2) \quad \hat{y}_{i,w} = e^{\mathbf{x}_i' \hat{\beta}_w},$$

où

$$\hat{\beta}_{wv} = \left( \sum_{i \in I} w_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \sum_{i \in I} w_i \mathbf{x}_i \log y_i \right).$$

Le modèle  $\mathbf{M}_2$  requiert une spécification subjective de la transformation appliquée à la variable étudiée, ce qui peut être raisonnable dans des situations où nous savons quelle est la transformation appropriée grâce à des données empiriques *a priori* ou à la théorie. En l'absence de toute connaissance *a priori* au sujet de la transformation, il est prudent de choisir cette dernière parmi une famille flexible de transformations en utilisant les données.

Tukey (1957) a examiné la famille de transformations puissance :

$$y^{(\lambda)} = \begin{cases} y^\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où  $y > 0$ . Afin d'éliminer la discontinuité à  $\lambda = 0$ , Box et Cox (1964) ont proposé la famille de transformations suivante :

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log(y) & \lambda = 0, \end{cases}$$

où  $y > 0$ . Le paramètre  $\lambda$  détermine la nature de la transformation. Par exemple, les valeurs  $\lambda = 1, 0, 0.5, -1$  correspondent à l'absence de transformation, une transformation logarithmique, une transformation racine carrée et une transformation inverse, respectivement. Le paramètre de transformation  $\lambda$  est estimé d'après les données. L'analyse de Box-Cox peut mener à une transformation logarithmique, mais tout aussi bien aboutir à une autre transformation de la famille susmentionnée, selon les données réelles observées.

Nous considérons le modèle de superpopulation suivant pour la variable étudiée transformée :

$$\mathbf{M}_3 : \mathbf{Y}^{(\lambda)} = \mathbf{X}\beta + \varepsilon,$$

Box-Cox peut être consultée dans un article de synthèse publié par Sakia (1992). Pour une application de la méthodologie de Box-Cox à un modèle linéaire mixte, voir Gurka (2004, 2006).

Li et Lahiri (2007) appliquent la transformation de Box-Cox à la variable étudiée pour obtenir des prédicteurs fondés sur un modèle d'un total de population finie qui sont robustes. Ils mentionnent aussi les estimateurs assistés par modèle dans une sous-section (section 2.6), mais n'étudient pas leurs propriétés. Le présent article fournit cette analyse. Nous employons la technique de Box-Cox pour ajuster une régression de la variable étudiée sur un ensemble de variables auxiliaires. Puis, nous utilisons la régression ajustée pour prédire les valeurs de la variable étudiée pour les unités non observées de la population finie qui, à leur tour, fournissent un estimateur de type régression adaptative dans le cadre de travail assisté par un modèle.

La présentation de l'article est la suivante. À la section 2, nous proposons un nouvel estimateur et examinons les propriétés analytiques de ce dernier en fonction du plan de sondage. À la section 3, pour mieux évaluer la robustesse et l'efficacité de l'estimateur proposé, nous le comparons aux estimateurs GREG fondés sur les modèles de travail linéaire et logarithme sous-jacents par simulations de Monte Carlo. À la section 4, nous discutons des résultats. Enfin, à la section 5, nous tirons certaines conclusions au sujet des domaines dans lesquels les travaux de recherche pourraient se poursuivre.

## 2. Un nouvel estimateur du total de population finie

Supposons que la quantité d'intérêt est le total de population finie

$$T = \sum_{i \in U} y_i,$$

où  $U = \{1, \dots, N\}$  dénote une population finie de taille connue  $N$ , et  $y_i > 0$  est la valeur de la variable étudiée associée à l'unité  $i$ . Écrivons  $\mathbf{Y} = (y_1, \dots, y_N)'$ . Pour estimer  $T$ , nous tirons un échantillon  $s$  de taille  $n$  de la population finie selon un plan d'échantillonnage probabiliste. Soit  $w_i$  les poids d'échantillonnage de l'unité  $i$ . Ce poids est simplement l'inverse de la probabilité d'inclusion de l'unité  $i$ , dénotée par  $\pi_i = P(i \in s) \ (i = 1, \dots, N)$ . Nous supposons que nous avons de l'information sur  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$ , où  $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})'$  est un vecteur colonne de  $k$  variables auxiliaires pour l'unité  $i$ . Pour tout échantillon  $s$  de taille  $n$ , nous redéfinissons  $\mathbf{Y}$  et  $\mathbf{X}$  de façon que les  $n$  premières lignes de  $\mathbf{Y}$  et  $\mathbf{X}$  correspondent à celles figurant dans l'échantillon. Écrivons

certaines applications des sondages, particulièrement les enquêtes auprès des entreprises et des établissements, les variables étudiées sont souvent continues, positives et fortement asymétriques (par exemple, le revenu). Pour estimer le total de population finie, l'application d'un modèle linéaire à la variable étudiée pourrait ne pas convenir, mais être raisonnable pour une transformation strictement monotone de cette variable. Chen et Chen (1996) ont envisagé l'utilisation de données d'enquête transformées pour améliorer la précision de l'approximation normale. Korn et Graubard (1998) ont comparé divers intervalles de confiance, y compris des intervalles fondés sur une transformation logit, pour l'estimation de proportions dans le cas d'un petit nombre prévu de dénombrements positifs. Karlberg (2000) a proposé un estimateur basé sur un modèle de superpopulation lognormal-logistique pour prédire le total de population finie d'une variable étudiée à distribution fortement asymétrique. Les résultats de simulation ont indiqué que l'estimateur fondé sur le modèle lognormal-logistique offre une alternative sensée à d'autres estimateurs, particulièrement quand la taille de l'échantillon est faible. Chambers et Dorfman (2003) ont discuté de l'estimation d'une moyenne de population finie sous certaines transformations générales, mais connues, des données continues.

Les chercheurs jugent la technique de transformation de données d'enquête. L'étape clé consiste à trouver une transformation appropriée, bien adaptée à ces données. Dans de nombreuses applications, la forme de la transformation est déterminée subjectivement. Malheureusement, les renseignements *a priori* ou la théorie ne suggèrent pas nécessairement celle qu'il convient d'utiliser. Le cas échéant, il serait commode d'identifier la transformation de manière adaptative, en utilisant les données d'enquête.

Les travaux de Box et Cox (1964) ont mené à l'élaboration de méthodes du type « les données décident de la transformation » pour la construction de modèles avec erreurs indépendantes et identiquement distribuées. Leur article et d'autres traitant du sujet, dont Turkey (1957), John et Draper (1980), ainsi que Bickel et Doksum (1981), ont examiné la relation entre la demande d'argent et la tappe à l'liquidité à l'aide d'un modèle de Box-Cox généralisé. Dans le contexte de la recherche sur la malaria, Newman (1977) a conclu que la spécification fonctionnelle de Box-Cox était supérieure aux spécifications antérieures. Miner (1982), ainsi que Davison, Arnade et Hallahan (1989) ont étudié la modélisation des fonctions du rendement du soja et de l'exportation de soja par les États-Unis, respectivement. Ils ont conclu que la transformation de Box-Cox produit des termes d'erreur qui suivent approximativement une loi normale. Une bibliographie sur la transformation de



# Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox

Yan Li<sup>1</sup>

Résumé

Un nouvel estimateur par la régression généralisée d'un total de population finie basé sur la méthode de transformation de Box-Cox et son estimateur de la variance sont proposés sous un plan général de sondage à probabilités inégales. En étant convergent par rapport au plan de sondage, l'estimateur proposé retient la propriété de robustesse de l'estimateur GREG, même si le modèle sous-jacent est déformé. En outre, la méthode de Box-Cox permet de trouver automatiquement une transformation raisonnable de la variable dépendante en se servant des données. La robustesse et l'efficacité du nouvel estimateur sont évaluées analytiquement et par des études en simulation de Monte Carlo.

Mots clés : Estimateur par la régression généralisée (GREG); méthode de Box-Cox; pseudo maximum de vraisemblance (PMV).

## 1. Introduction

Les estimateurs par la régression généralisée (GREG) de totaux et de moyennes de population finie sont établis au moyen de modèles de régression appropriés. Bien que l'on se serve de modèles pour construire ces estimateurs, la sélection des échantillons et l'évaluation des propriétés statistiques des stratégies d'estimation résultantes doivent se faire sous randomisation. Des exemples peuvent être consultés dans Särndal, Swensson et Wretman (1992), Estévez, Hidroglou et Särndal (1995), Fuller, Louglin et Baker (1994), ainsi que Jayasunya et Valliant (1996). Le choix d'un bon modèle est essentiel si l'on veut limiter la variabilité d'un estimateur assisté par modèle tel que l'estimateur GREG. Si le modèle hypothétique décrit bien la population finie, l'utilisation de l'estimateur GREG au lieu de l'estimateur d'Horvitz-Thompson (Horvitz et Thompson 1952) peut réduire considérablement la variance. Une discussion générale de l'estimation par la régression est présentée dans Fuller (2002). Särndal et coll. (1992) offrent une description détaillée du cadre de travail assisté par modèle pour la construction d'estimateurs de sondage. Les études portant sur l'estimateur GREG ont été réalisées pour la plupart dans le contexte de la modélisation par la régression linéaire. Essentiellement, l'estimateur GREG intègre les variables auxiliaires pertinentes par la voie de leurs totaux de population de contrôle connus, même si les variables auxiliaires sont connues pour chaque unité de la population (Cassel, Särndal et Wretman 1976; Särndal, Deville et Särndal 1992; Särndal et coll. 1992; Jiang et Lahiri 2006). De nos jours, il est assez fréquent que l'on dispose d'information auxiliaire complémentaire : les données de recensement, les registres administratifs, les données de télédétection et les enquêtes antérieures

constituent une mine de renseignements précieux qui peuvent être utilisés pour accroître la précision de la procédure d'estimation (Montanari et Ranalli 2003). Par conséquent, des modèles complexes et des méthodes souples s'appuyant sur l'information auxiliaire complète ont été introduits dans les sondages ces dernières années. Les techniques des splines pénalisées ont été adaptées pour construire des estimateurs fondés sur un modèle (Zhang et Little 2004) et assistés par un modèle (Bridt, Claeskens et Opsomer 2005) pour un total de population finie en se basant sur des données d'enquête complexe. Bridt et Opsomer (2000) ont considéré un estimateur par la régression non paramétrique, assisté par modèle, recourant à la régression par polynômes locaux et ont montré que la régression non paramétrique peut accroître significativement l'efficacité des estimateurs quand les modèles paramétriques sont spécifiés incorrectement. Leurs travaux ont été étendus du modèle à une seule covariable au cas du modèle additif semi-paramétrique. Wu et Sitter (2001) ajustent un modèle de travail général, qui pourrait comprendre à la fois des composantes linéaires et non linéaires, puis le calent sur les valeurs prédites résultantes par régression linéaire simple. Montanari et Ranalli (2005) combinent l'estimation par calage fondé sur un modèle et des méthodes non paramétriques, et proposent des estimateurs non paramétriques assistés par modèle pour une moyenne de population finie. Alors qu'en statistique classique, on procède souvent à une transformation appropriée de la variable dépendante dans le modèle hypothétique pour obtenir la normalité, la littérature sur les transformations dans le contexte de l'inférence en population finie n'est pas très riche. L'intérêt croît cependant pour l'élaboration de méthodes utilisant une transformation appropriée et des données de sondage. Dans

Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. Dans *Bayesian Statistics 2*, (Eds., J.M. Bernardo, M.H. DeGroot, D.V. Lindley et A.F.M. Smith). Amsterdam : Elsevier/North-Holland, 463-472.

Sæmdal, C.-E., Swensson, B. et Wretnan, J. (1992). *Model Assisted Survey Sampling*. New York : Springer.

Sasry, N., Ghosh-Dastidar, B., Adams, J. et Pebley, A.R. (2005). The design of a multilevel survey of children, families, and communities: The Los Angeles Family and Neighborhood Survey. *Social Science Research*, 35, 1000-1024.

Thomsen, I., Tesfu, D. et Binder, D.A. (1986). Estimation of design effects and intraclass correlations when using outdated measures of size. *Revue Internationale de Statistique*, 54, 343-349.

Valiant, R., Dorfman, A.H. et Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York : John Wiley & Sons, Inc.

Waksberg, J. (1973). The effect of stratification with differential sampling rates on attributes of subsets of the population. Dans *ASA Proceedings of the Social Statistics Section*, American Statistical Association. Alexandria, VA.

Waksberg, J. (1995). Distribution of poverty in census block groups (BGs) and implications for sample design. Dans *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association. Alexandria, VA, 497-502.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.

Denton, N.A., et Massey, D.S. (1993). *American Apartheid: Segregation and the Making of the Underclass*. Cambridge, MA : Harvard University Press.

Dumouchel, W.H., et Duncan, G.J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

Eaton, D.K., Kann, L., Kinchen, S., Ross, J., Hawkins, J., Hartts, W.A., Lowry, R., Mcmanus, T., Chyen, D., Shanklin, S., Lim, C., Grunbaum, J.A. et Wechsler, H. (2006). Youth risk behavior surveillance - United States, 2005. *Morbidity and Mortality Weekly Report*, 55(SS-5), 1-108.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sanhitya*, C37, 117-132.

Fuller, W.A. (1984). Application de la méthode des moindres carrées et de techniques connexes aux plans de sondage complexes. *Techniques d'enquête*, 10, 107-130.

Fuller, W.A. (2002). Estimation par régression appliquée à l'échantillonnage. *Techniques d'enquête*, 28, 5-25.

Hansen, M.H., Madow, W.G. et Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78, 776-793.

Isaki, C.T., et Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-97.

Kish, L. (1974). Optimal and proximal multipurpose allocation. Alexandria, VA. In *Proceedings of the Social Statistics Section, American Statistical Association*, 111-118.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

McCullagh, P., et Nelder, J.A. (1989). *Generalized Linear Models* (Deuxième édition). London : Chapman & Hall Ltd.

Morel, J.G. (1989). Régression logistique selon des plans de sondage complexes. *Techniques d'enquête*, 15, 213-233.

Morris, C., Newhouse, J.P. et Archibald, R. (1979). On the theory and practice of obtaining unbiased and efficient samples in social surveys and experiments. *Research in Experimental Economics*, 1, 199-220.

Nathan, G., et Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society, Series B, Methodological*, 42, 377-386.

Neyman, J. (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

## Remerciements

La présente étude a été financée par les subventions U01-CA93344 (Zaslavsky et Zheng), U01-CA93324 (Zaslavsky) et U01-CA093348 (Adams) du National Cancer Institute. Les auteurs remercient Nat Schenker et Van Parsons de leurs commentaires constructifs concernant une version antérieure, ainsi que le rédacteur adjoint et deux examinateurs de leurs commentaires judicieux.

## Bibliographie

Ayamian, J.Z., Chntsilles, E.A., Wallace, R.B., Fletcher, R.H., Fouad, M., Kiefe, C.J., Harrington, D.P., Weeks, J.C., Kahn, K.L., Malin, J.L., Lipscomb, J., Polosky, A.L., Provenzale, D.T., Sandler, R.S., Vanny, M. et West, D.W. (2004). Understanding cancer treatment and outcomes: The Cancer Outcomes Research and Surveillance Consortium. *Journal of Clinical Oncology*, 2, 2992-2996.

Bellhouse, D.R. (1984). A review of optimal designs in survey sampling. *The Canadian Journal of Statistics*, 12, 53-65.

Binder, D.A. (1981). Les variances d'estimateurs asymptotiquement normaux bases sur des enquetes complexes. *Techniques d'enquête*, 7, 162-178.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

Botman, S.T., Moore, T.F., Moriarty, C.N. et Parsons, V.T. (2000). *Design and Estimation for the National Health Interview Survey, 1995-2004*. Vital and Health Statistics, 2(130). Washington, DC : National Center for Health Statistics.

Cassel, C.M., Särndal, C.-E. et Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chromy, J.R. (1987). Design optimization with multiple objectives. *Dans Proceedings of the Section on Survey Research Methods, American Statistical Association*. Alexandria, VA. 194-199.



études énumératives suscite relativement peu de controverses, leur utilisation dans les études analytiques a provoqué un vif débat (Hansen, Madow et Tepping, 1983 et discussion; DuMouchel et Duncan 1983; Bellhouse 1984; Pfeffermann 1993; Fuller 2002, section 5). Une analyse descriptive de la population offre une certaine robustesse à la sélection éventuelle de l'échantillon d'une manière qui fausse les relations typiques. Donc, même si l'on ne peut pas justifier une analyse GDP pure pour des raisons de représentativité énumérative, un échantillon tiré pour optimiser l'estimation non pondérée des coefficients de régression pourrait avoir une valeur scientifique limitée. Ainsi, supposons que les données CanCORS soient analysées par une régression *non pondérée* pour estimer un simple effet du revenu (un contraste des moyennes) en utilisant l'information sur les variables du plan d'échantillonnage au niveau de l'ilot provenant du recensement. De manière optimale, l'échantillon serait tiré à partir d'un ensemble d'ilot pour lesquels, collectivement, environ la moitié des résidents vivraient dans la pauvreté. Puisque les taux de pauvreté sont rarement aussi élevés, cette situation requiert effectivement un échantillonnage uniquement dans les ilots où les taux de pauvreté sont les plus élevés. Pareil échantillonnage ne serait représentatif d'aucune des catégories de revenu. De même, un échantillon qui sur-représente les résidents noirs par échantillonnage d'ilot sans pondération) non représentatif de la population noire en général, parce que les services offerts dans les régions où la ségrégation est forte diffèrent vraisemblablement de ceux offerts dans des régions mixtes.

Il est nécessaire d'établir des formules plus générales, avec les hypothèses et les objectifs clairement énoncés, dans lesquelles « les paramètres du modèle sont considérés comme les paramètres cibles ultimes, mais se concentrent en même temps sur les grands traits descriptifs de la population (GDP) comme moyen de rendre l'inférence robuste » (Pfeffermann 1993), en tenant compte des objectifs scientifiques de l'étude. Il a été proposé antérieurement de tester l'hypothèse nulle que les poids n'ont pas d'effet sur la régression (DuMouchel et Duncan 1983; Fuller 1984) ou d'inclure les variables du plan d'échantillonnage (Nathan et Holt, 1980; Little 1991) ou les poids proprement dits (Rubin 1985) comme variables de contrôle dans la régression. Ces approches posent néanmoins des problèmes quand les poids sont des fonctions des covariables d'intérêt scientifique primordial. Nous avons essayé, grâce à la pondération flexible des contrastes (section 2.5), de faire un pas vers une formule générale, en étendant l'approche GDP afin de mettre l'accent sur une gamme d'inférences valides pour des objectifs scientifiques particuliers, plutôt qu'exclusivement sur l'inférence pour des populations finies. À partir de cette

gamme, un chercheur peut choisir un objectif inférentiel et un plan d'échantillonnage adapté à la structure de la population et aux aspects pratiques de la conception de l'étude.

## Annexe

### Équivalence des variances sous le plan d'échantillonnage et sous le modèle

Nous montrons que  $E_{\xi} \text{Var}_{\pi} \beta \approx E_{\pi} \text{Var}_{\xi} \beta$  sous les contraintes suivantes :

1.  $1/N \mathbf{X}'_U \mathbf{X}_U \rightarrow \Sigma$  pour une certaine  $\Sigma$  définie positive. Cette condition minimale relie la série hypothétique de populations.

2. L'estimateur par la régression sous le plan de sondage peut s'écrire  $\beta = \beta_U + R_{n_s}$ , où  $\text{Var}_{\pi} E_{\pi} R_{n_s} = o(n^{-1})$  et  $\text{Var}_{\pi} E_{\xi} R_{n_s} = o(n^{-1})$ . Notons que  $\beta$  ne peut pas être défini strictement comme dans (2), parce que la matrice inverse est indéfinie quand les valeurs d'échantillon de  $x$  ne couvrent pas l'espace du plan d'échantillonnage et que son espérance et sa variance sont également indéfinies. Un estimateur par le ratio scalaire pourrait de même être indéfini avec une probabilité non nulle mais à l'ordre  $o(n^{-1})$ , parce que l'échantillon pourrait n'avoir que des valeurs 0 pour la variable du dénominateur. Si nous attribuons une valeur arbitraire à cet événement, l'estimateur pourrait néanmoins avoir de bonnes propriétés asymptotiques. Un argument semblable nous permet de supposer qu'un  $\beta$  approprié peut être défini. Nous ne spécifions pas comment (2) doit être modifié pour satisfaire techniquement la contrainte, puisque celle-ci dépend des caractéristiques particulières de  $\xi$  et de la série de plans d'échantillonnage.

3.  $\max(\pi_j) = O(n/N)$  et  $n = o(N)$ , essentiellement l'omage.
4. Homoscédasticité,  $\text{Var}_{\xi} y_i = \sigma^2$ , cela n'est pas finie peuvent être ignorées.

Homoscédasticité,  $\text{Var}_{\xi} y_i = \sigma^2$ , cela n'est pas finie peuvent être ignorées.

Sous ces contraintes,

$$\begin{aligned} \text{Var}_{\pi_{\xi}}(\beta) &= \text{Var}_{\pi} E_{\xi} \beta + E_{\pi} \text{Var}_{\xi} \beta \\ &= \text{Var}_{\pi} \beta + E_{\pi} \text{Var}_{\xi} \beta \\ &= o(n^{-1}) + E_{\pi} \text{Var}_{\xi} \beta \end{aligned}$$

statistique d'influence des cas et, par conséquent, réduire la variance des estimations des coefficients (Fulter, 1975). Puisque l'efficacité du plan est insensible à de faibles écarts par rapport à la situation optimale, certains plans stratifiés avec probabilités égales dans les strates pourraient approcher l'efficacité du plan optimal. Des stratifications *ponctuelles* pourraient être moins efficaces, même sous stratification des ilots en fonction du groupe race-revenu représenté dont la prévalence est la plus faible produit un plan d'échantillonnage dont le gain d'efficacité est égal à environ la moitié de celui de notre plan d'échantillonnage avec utilisation de données agrégées au niveau de l'ilot.

En ce qui concerne le dernier point, il convient de souligner que la conception de strates homogènes pour l'estimation des coefficients de régression est vraisemblablement plus difficile que pour l'estimation d'une moyenne. L'influence d'une observation dépend du résidu de cette dernière produit par le modèle de régression et non de sa valeur brute, de sorte que pour réduire l'homogénéité, la stratification devrait comprendre des variables prédictives qui ne sont pas incluses dans le modèle. L'influence dépend aussi de l'effet levier de l'observation pour chaque coefficient, fait qui pourrait être une fonction complexe des covariables.

Dans le cas de l'échantillonnage par grappe, l'équivalence de  $E_{\text{Var}} \hat{\alpha} \hat{\beta}$  et  $E_{\text{Var}} \hat{\alpha} \hat{\beta}$  pourrait ne pas tenir, sauf sous des hypothèses contraignantes, telles que l'indépendance des résidus; donc, les termes du facteur du milieu de (5) prendraient une forme plus complexe. Plusieurs cas sont possibles pour l'échantillonnage par grappe, selon la relation entre les cellules et les grappes, et il conviendrait de les élaborer dans le cadre de futurs travaux.

Une autre extension naturelle est celle aux modèles de régression non linéaire et à d'autres grandsurcs à estimer définies par des équations d'estimation. La formulation des moindres carrés pondérés de l'étape de Newton-Raphson (McCullagh et Nelder 1989, section 2.5) pour un modèle linéaire généralisé peut être appliquée en définissant de manière appropriée  $\sigma^2_{\hat{\alpha}}$  dans (3) et donc dans (4) à (6); une procédure semblable peut être appliquée à d'autres équations d'estimation (Binder 1981; Binder 1983; Morel 1989). Comme les variantes sont des fonctions des prédictions du modèle, la mise en œuvre de ces modifications nécessite des hypothèses de conception du plan d'échantillonnage concernant le modèle ajusté, ainsi que la distribution des covariables.

Toute optimisation a son prix, lequel, dans le cas de nos méthodes, peut être pratique et statistique. Dans l'étude CanCORS, les nouveaux cas de cancer étudiés ont été identifiés en temps réel grâce à une opération sur le terrain (« confirmation rapide des cas »); une prise de

## D'échantillonnage problématique.

Ce genre de situation pose moins de problèmes dans les sondages à l'aide d'une base de sondage statique qui peut être traitée dans des délais moins contraignants, particulièrement dans les sondages à grande échelle et (ou) répétés dans lesquels même une réduction modeste de la variance justifie une complexité supplémentaire. Ils pourraient être utilisés, par exemple, pour évaluer les améliorations éventuelles grâce au surechantillonnage à base géographique dans les sondages pour lesquels des estimations nationales selon la

race sont requises.

Les préoccupations statistiques concernant d'autres stratégies d'élaboration du plan d'échantillonnage découlent du fait que l'optimisation pour un ensemble d'objectifs statistiques prédéterminés réduira vraisemblablement l'efficacité pour d'autres ensembles. À part pour les études dont les objectifs sont cernés le plus étroitement, il est généralement difficile de prévoir toutes les analyses possibles. L'optimisation simultanée pour un ensemble raisonnablement complet d'analyses, et l'étude de la sensibilité du plan d'échantillonnage à la variation des poids relatifs des divers objectifs devrait offrir une certaine protection contre un plan d'échantillonnage surspécialisé. Cependant, cette approche ne peut être utilisée qu'avec des variables pour lesquelles il existe certaines données avant l'étude. Les résultats de la section 2.6 donnent à penser que la surveillance de l'effet de l'échantillonnage non proportionnel sur la précision de la moyenne de population offre une certaine protection contre les plans d'échantillonnage excessivement inefficaces pour les analyses et les variables non prévues, quoique les bornes ne soient pas très générales. De manière plus générale, nous pourrions nous demander quand l'analyse GDP correspond à la grandeur à estimer scientifiquement pertinente. Les modèles de régression sont souvent utilisés dans des analyses destinées à être généralisables à des populations plus générales, plutôt qu'à décrire la population finie dont on dispose, tout comme les emplacements de l'étude CanCORS ont été choisis à dessein pour étudier les profils et les variations des soins qui pourraient refléter les tendances nationales plus générales. Bien que l'utilisation de poids d'échantillonnage dans des



**Tableau 3**  
Variances normalisées et poids des contrastes pour des plans d'échantillonnage GDP optimaux avec pondération flexible des contrastes du revenu selon la race. Des lignes pour les contrastes fixes sont incluses afin de démontrer l'effet de divers choix de pondération flexible, en vue de la comparaison à des scénarios d'objectif avec pondération fixe.

Variances pour l'estimation « sous pondération flexible » et pour les contrastes présentes au tableau 2									
Variance en pourcentage de la variance sous EAS									
(selon l'information disponible sur les variables du plan)									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									
Race+									



I'hypothèse d'additivité de ce modèle. Par ailleurs, la façon dont les races sont combinées ne reflète pas nécessairement les proportions dans la population. Les poids appliqués au contraste du revenu dans chaque race, estimés comme nous l'avons décrit à la section 2.5, sont présentés à la partie inférieure du tableau 3 pour démontrer comment cette approche nous permet de modifier la grandeur estimée afin d'exploiter l'information disponible sur les variables du plan. (L'algorithme d'optimisation alternée a convergé vers une précision adéquate en sept itérations.)

Sous EAS, la variance de l'effet du revenu sous le modèle à pondération flexible est un peu plus grande que sous le modèle à deux facteurs (1,5,91 contre 14,99). Le poids accordé au contraste de la race blanche sous ce plan (51 %) est plus faible que la proportion de Blancs dans la population (66 %), parce que relativement peu de Blancs sont pauvres et que, par conséquent, le contraste du revenu parmi les Blancs est relativement imprécis. Inversement, le poids appliqué au contraste du revenu chez les Noirs (12 %) est presque le double de la part de ce groupe dans la population, à cause du taux de disproportionnellement élevé de pauvreté dans ce groupe.

L'utilisation de données sur la race de niveau individuel dans le plan d'échantillonnage accentue cette disproportion, une plus grande part de l'échantillon et une plus forte pondération (75 %) étant données au groupe des Noirs, avec le pourcentage le plus élevé de pauvreté. Donc, la pondération flexible permet de réduire considérablement la variance de l'effet estimé du revenu (à 63 % de sa valeur sous EAS) en utilisant uniquement la race, ce qui était impossible sous le modèle GDP à deux facteurs plus restrictif.

L'information au niveau de l'ilot est un peu moins utile pour ce plan d'échantillonnage que celle sur la race. La combinaison de l'information sur l'ilot et sur la race est toutefois très puissante, réduisant la variance de l'effet du revenu à 48 % de sa valeur sous EAS. Sous ce plan d'échantillonnage, beaucoup plus de poids (46 %) est accordé au contraste du revenu des Hispaniques, qui peut être estimé efficacement à cause de la ségrégation plus forte du revenu chez ce groupe (tableau 1). Si l'on dispose de données individuelles sur le revenu (avec ou sans la race), les poids des contrastes correspondent approximativement aux proportions par race, puisque des contrastes de revenu efficaces peuvent être obtenus pour toute race et que l'inclusion de la moyenne globale dans l'objectif déplace le plan d'échantillonnage vers l'échantillonnage proportionnel. Donc, le plan d'échantillonnage diffère radicalement sous diverses hypothèses concernant la disponibilité de l'information sur les variables du plan.

Au tableau 4, nous comparons les améliorations pour l'échantillonnage non proportionnel à quatre emplacements géographiques et un emplacement (Iowa) dont la population était presque entièrement blanche. Pour chaque emplacement, nous avons procédé à l'optimisation, pour la pondération unitaire ( $p_i = 1$ ), des variances de la moyenne globale et de l'effet du revenu dans le modèle à deux facteurs (proportionnel à l'objectif 7), sous diverses hypothèses concernant l'information disponible sur les variables du plan d'échantillonnage. Le minimum théorique pour cet objectif avec une population équilibrée est 5 ( $V_{\text{revenu}}^{\text{moyenne}} = 1$ ). L'EAS est inefficace à chaque emplacement, surtout en Alabama et dans le Nord de la Californie, et l'information sur la race serait, à elle seule, de peu d'utilité. Inversement, la meilleure variance qui peut être obtenue en utilisant l'information complète sur la race et sur le revenu au niveau individuel est comprise entre 5,60 et 5,72 à chaque emplacement. Le suréchantillonnage fondé sur l'information sur le revenu au niveau de l'ilot réduirait considérablement les variances, les progrès étant nettement plus importants en Alabama et dans le Nord de la Californie que pour les autres emplacements.

#### 4. Discussion

Afin d'élaborer divers plans d'échantillonnage pour une étude sur les services de santé, nous avons étendu les méthodes antérieures d'élaboration de plans optimaux pour l'estimation par domaine afin de montrer comment un plan d'échantillonnage avec probabilités inégales optimal doit être conçu pour estimer les coefficients de régression d'un ou de plusieurs modèles. Dans notre application, nous avons pu réaliser des réductions importantes de la variance même quand l'information sur certaines variables n'était disponible que pour des agrégats géographiques. Nous avons obtenu des améliorations particulièrement importantes pour les variables explicatives catégoriques (état de pauvreté, race) présentant des distributions très asymétriques. Essentiellement, notre approche d'établissement du plan d'échantillonnage lorsque les variables du plan sont mesurées des variables de manière imprécise s'appuie sur la distribution prédictive des variables du plan pour chaque unité échantillonnée, plus précisément les espérances des variables, ainsi que leur carré et leur produit croisé. Ce concept unit le plan d'échantillonnage utilisant des agrégats de cellules (estimates d'après des données de recensement ou d'échantillon), utilisant des variables mesurées avec erreur, ou utilisant une base de sondage dont les caractéristiques des unités pourraient avoir changé au cours du temps.

Tableau 2 (suite)  
Variances normalisées et fonctions objectif pour les plans d'échantillonnage optimaux pour divers poids objectifs et hypothèses concernant l'information disponible sur les variables du plan

Objectif 5 : Optimisation pour des effets de race distincts, l'effet du revenu et la moyenne globale												
Variances en pourcentage de la variance sous EAS												
(selon l'information disponible sur les variables du plan)												
Variance												
Modèle	Effet	(d <sub>i</sub> )	EAS	Race	l'lot	Race+	Revenu	Race+	Revenu	Race+	Revenu	Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>
Constante	Moyenne	3	1,0	111	117	135	114	109	112	112	116	53
Race	Noirs	1	17,2	54	74	55	109	109	112	112	116	67
Race	Asiatiques	1	7,5	95	106	109	112	112	112	112	116	67
Revenu	Pauvres	3	15,0	101	82	72	38	109	112	112	116	37
Race+Revenu	Noirs	0	17,4	55	74	55	109	109	112	112	116	52
Race+Revenu	Asiatiques	0	7,5	95	106	109	112	112	112	112	116	115
Race+Revenu	Hispaniques	0	13,8	67	96	69	112	112	112	112	116	66
Race+Revenu	Pauvres	0	15,2	101	82	72	39	112	112	112	116	35
Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>	Pauvres	0	86,4	86	86	74	73	86	86	86	86	56

Objectif 6 : Optimisation pour les effets de race et l'effet du revenu dans un modèle à deux facteurs et pour la moyenne globale												
Variances en pourcentage de la variance sous EAS												
(selon l'information disponible sur les variables du plan)												
Variance												
Modèle	Effet	(d <sub>i</sub> )	EAS	Race	l'lot	Race+	Revenu	Race+	Revenu	Race+	Revenu	Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>
Constante	Moyenne	3	1,0	111	119	138	114	108	112	112	114	156
Race	Noirs	0	17,2	55	74	55	108	108	112	112	114	53
Race	Asiatiques	0	7,5	95	107	109	112	112	112	112	114	67
Revenu	Hispaniques	0	13,7	67	96	69	111	111	112	112	114	67
Revenu	Pauvres	0	15,0	101	82	72	38	109	112	112	114	37
Race+Revenu	Noirs	1	17,4	55	74	55	109	109	112	112	114	52
Race+Revenu	Asiatiques	1	7,5	95	107	109	113	113	113	113	113	113
Race+Revenu	Hispaniques	1	13,8	67	96	69	112	112	112	112	113	66
Race+Revenu	Pauvres	3	15,2	101	81	71	39	112	112	112	113	35
Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>	Pauvres	0	87,2	86	86	73	73	86	86	86	86	54

Objectif 7 : Optimisation pour l'effet du revenu dans un modèle à deux facteurs et pour la moyenne globale												
Variances en pourcentage de la variance sous EAS												
(selon l'information disponible sur les variables du plan)												
Variance												
Modèle	Effet	(d <sub>i</sub> )	EAS	Race	l'lot	Race+	Revenu	Race+	Revenu	Race+	Revenu	Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>
Constante	Moyenne	3	1,0	103	135	149	148	139	145	144	156	122
Race	Noirs	0	17,2	77	100	98	148	139	145	144	156	97
Race	Asiatiques	0	7,5	91	124	132	145	132	145	144	156	132
Revenu	Hispaniques	0	13,7	86	122	135	144	132	145	144	156	122
Revenu	Pauvres	0	15,0	97	75	62	28	132	145	144	156	28
Race+Revenu	Noirs	0	17,4	77	100	98	140	132	145	144	156	96
Race+Revenu	Asiatiques	0	7,5	91	124	132	146	132	145	144	156	132
Race+Revenu	Hispaniques	0	13,8	86	122	135	146	132	145	144	156	121
Race+Revenu	Pauvres	3	15,2	97	75	62	29	132	145	144	156	27
Total = nV = nΣd <sub>i</sub> V <sub>a<sub>i</sub></sub>	Pauvres	0	48,6	97	79	67	36	132	145	144	156	35

3.4 Efficacité sous des poids de contraste flexibles

Nous examinons maintenant les avantages éventuels de l'estimation des effets du revenu sous un scénario de pondération flexible (tableau 3). La fonction objectif prend en considération les coefficients de deux modèles, le modèle constant dont le paramètre est la moyenne de population et un modèle contenant des variables indicatrices pour chaque cellule race par revenu. Nous estimons l'effet du revenu à

l'intérieur de chaque race comme étant la différence entre les coefficients pour les pauvres et les non-pauvres à l'intérieur de la race, puis nous combinons ces estimations avec des poids flexibles pour estimer un effet global du revenu. Cette stratégie s'approche plus de l'objectif 7, qui comporte aussi l'estimation des effets du revenu en tenant compte de l'effet de la race. L'analyse des contrastes souples dépend moins du modèle que l'option du modèle à deux facteurs en ce sens qu'elle ne s'appuie pas sur

**Tableau 2**  
**Variances normalisées et fonctions objectif pour les plans d'échantillonnage optimaux pour divers poids objectifs et hypothèses concernant l'information disponible sur les variables du plan**

Objectif 1 : Optimisation pour les effets de race											
Variances en pourcentage de la variance sous EAS											
(selon l'information disponible sur les variables du plan)											
Modèle	Effet	(d <sub>i</sub> )	Pondération	Variances	Race	Illot	Illot	Race+	Revenu	Race+	Revenu
Constante	Moyenne	0,1	1,0	181	119	181	100	100	181	100	181
Race	Noirs	1	17,2	43	65	43	100	99	43	100	43
Race	Asiatiques	1	7,5	106	106	100	100	99	100	100	100
Race	Hispaniques	1	13,7	90	90	55	100	100	55	100	55
Revenu	Pauvres	0	15,0	182	104	182	182	81	182	100	182
Race+Revenu	Noirs	0	17,4	44	65	44	99	99	44	100	44
Race+Revenu	Asiatiques	0	7,5	100	106	100	100	100	100	100	100
Race+Revenu	Hispaniques	0	13,8	55	90	55	100	100	55	100	55
Pauvres	Pauvres	0	15,2	182	104	182	182	81	182	100	182
Total = $nV = n \sum d_i V_i$			38,6	59	82	59	59	99	59	99	59
Objectif 2 : Optimisation pour les effets de race et la moyenne globale											
Variances en pourcentage de la variance sous EAS											
(selon l'information disponible sur les variables du plan)											
Modèle	Effet	(d <sub>i</sub> )	Pondération	Variances	Race	Illot	Illot	Race+	Revenu	Race+	Revenu
Constante	Moyenne	3	1,0	136	115	136	100	100	136	100	136
Race	Noirs	1	17,2	44	66	44	99	99	44	100	44
Race	Asiatiques	1	7,5	100	104	100	100	100	100	100	100
Race	Hispaniques	1	13,7	90	90	56	100	100	56	100	56
Revenu	Pauvres	0	15,0	121	101	121	82	82	121	100	121
Race+Revenu	Noirs	0	17,4	45	66	45	99	99	45	100	45
Race+Revenu	Asiatiques	0	7,5	100	104	100	100	100	100	100	100
Race+Revenu	Hispaniques	0	13,8	56	90	56	100	100	56	100	56
Pauvres	Pauvres	0	15,2	122	102	122	82	82	122	100	122
Total = $nV = n \sum d_i V_i$			41,5	65	84	65	65	100	65	100	65
Objectif 3 : Optimisation pour les effets du revenu											
Variances en pourcentage de la variance sous EAS											
(selon l'information disponible sur les variables du plan)											
Modèle	Effet	(d <sub>i</sub> )	Pondération	Variances	Race	Illot	Illot	Race+	Revenu	Race+	Revenu
Constante	Moyenne	0,001	1,0	103	154	173	173	173	103	173	103
Race	Noirs	0	17,2	75	119	152	163	163	75	152	163
Race	Asiatiques	0	7,5	90	144	170	173	173	90	170	173
Race	Hispaniques	0	13,7	86	142	196	168	168	86	196	168
Revenu	Pauvres	0	15,0	97	74	60	27	27	97	60	27
Race+Revenu	Noirs	0	17,4	75	119	153	164	164	75	153	164
Race+Revenu	Asiatiques	0	7,5	90	144	174	171	171	90	174	171
Race+Revenu	Hispaniques	0	13,8	86	143	197	169	169	86	197	169
Pauvres	Pauvres	0	15,2	97	75	63	29	29	97	63	29
Total = $nV = n \sum d_i V_i$			45,0	97	74	60	27	27	97	60	27
Objectif 4 : Optimisation pour les effets du revenu et la moyenne globale											
Variances en pourcentage de la variance sous EAS											
(selon l'information disponible sur les variables du plan)											
Modèle	Effet	(d <sub>i</sub> )	Pondération	Variances	Race	Illot	Illot	Race+	Revenu	Race+	Revenu
Constante	Moyenne	3	1,0	103	134	147	151	151	103	147	151
Race	Noirs	0	17,2	76	107	128	142	142	76	128	142
Race	Asiatiques	0	7,5	91	127	145	148	148	91	145	148
Race	Hispaniques	0	13,7	86	125	161	147	147	86	161	147
Revenu	Pauvres	3	15,0	97	75	61	27	27	97	61	27
Race+Revenu	Noirs	0	17,4	76	107	129	143	143	76	129	143
Race+Revenu	Asiatiques	0	7,5	91	127	146	149	149	91	146	149
Race+Revenu	Hispaniques	0	13,8	86	125	162	147	147	86	162	147
Pauvres	Pauvres	0	15,2	97	75	63	29	29	97	63	29
Total = $nV = n \sum d_i V_i$			48,0	97	79	66	35	35	97	66	35



3.3 Efficacité sous modèles fixes

effets de race augmentent considérablement sous ces plans d'échantillonnage, mais l'accroissement de la pondération de la moyenne (objectif 4) donne des résultats nettement meilleurs en ce qui concerne l'amélioration de la variance de la moyenne et des effets de race, et n'accroît que légèrement la variance de l'effet estimé du revenu.

L'introduction des effets de la race ainsi que du revenu dans l'objectif 5 produit des plans d'échantillonnage qui ne sont pas aussi bons que les plans optimaux pour les uns et les autres de ces effets pris individuellement, mais malgré tout nettement meilleurs que sous EAS. Par exemple, les variances des effets de race lorsque l'on connaît la race et l'lot de résidence sont de 10 % à 24 % plus élevées que sous les plans d'échantillonnage utilisant la même information sur les variables du plan, mais optimisés séparément pour la race ou le revenu. Si l'on connaît uniquement l'information individuelle sur la race ou sur le revenu, le plan est essentiellement optimisé pour les effets correspondants à la variable disponible, en accroissant la variance des effets estimés de l'autre variable.

Le plan d'échantillonnage optimisé pour les effets conjugués de la race et du revenu dans le modèle additif à deux facteurs (objectif 6) donne des résultats assez proches de l'optimisation pour les effets de race et de revenu dans des modèles marginaux distincts (objectif 5). Quand nous optimisons des effets distincts, les variances de ces derniers sont un peu plus faibles que celles obtenues pour les effets correspondants dans le modèle à deux facteurs. Quand nous optimisons les effets dans le modèle conjoint, leurs variances sont réduites, mais, dans la plupart des cas, demeurent un peu plus élevées que celles des effets correspondants dans les modèles marginaux de race et de revenu, à cause de la confusion partielle des effets de race et de revenu.

De même, l'optimisation de l'effet du revenu dans le modèle à deux facteurs (objectif 7) est assez semblable à l'optimisation de l'effet univarié du revenu (objectif 4) lorsqu'on ne dispose d'aucune donnée sur la race. Cependant, la disponibilité de données sur la race, ainsi que sur le revenu au niveau de l'lot ou de l'individu réduit considérablement les variances des effets de race sous le plan d'échantillonnage pour le modèle à deux facteurs. À cause de la confusion partielle des effets de la race et du revenu sous ce modèle, ce plan d'échantillonnage est adapté afin d'estimer les premiers plus efficacement, en accumulant plus de données aux points du plan d'échantillonnage qui sont essentiels à l'élimination de la confusion entre ces effets.

Les deux premières fonctions objectives optimisent

l'estimation des contrastes de race et de la moyenne globale. L'utilisation de données individuelles sur la race améliore considérablement l'efficacité de l'estimation des effets des races noires et hispaniques. Les améliorations les plus importantes sont observées pour l'effet de la race noire (la plus petite des trois grandes minorités raciales, dont la variance est réduite à 43 % de sa valeur sous EAS. Inversement, nous n'obtenons aucune amélioration pour les Asiatiques-américains, dont la représentation dans la population est proche du taux d'échantillonnage optimal. Dans cette fonction objectif, une fois que l'information sur la race est disponible, des renseignements supplémentaires sur les variables du plan (revenu au niveau de l'lot ou de l'individu) n'a aucune importance pour l'optimisation. Si la race au niveau individuel est inconnue, l'utilisation de l'lot de résidence peut faciliter le suréchantillonnage des Noirs (le groupe pour lequel la ségrégation résidentielle est la plus importante), ce qui réduit la variance de l'effet estimé de la race noire à environ 65 % de sa valeur sous EAS, mais le suréchantillonnage par lot uniquement ne réduit que légèrement la variance de l'effet estimé de la race hispanique. Connaître le revenu est, en soi, de peu d'utilité en ce qui concerne l'amélioration de l'échantillonnage pour l'esti-

mation des effets de race. L'échantillonnage non proportionnel, ajusté pour optimiser l'estimation des effets de la race, accroît les variances des autres estimateurs de paramètre. Si l'on applique un poids minimal à la moyenne dans l'objectif d'optimisation (objectif 1), cette augmentation peut être assez importante, soit un facteur de 181 % pour la moyenne et les effets du revenu. Accorder plus de poids à la moyenne (objectif 2) modère cet effet, réduisant l'accroissement de la variance à 136 % pour la moyenne et 121 % pour l'effet du revenu, tout en n'accroissant que légèrement les variances pour les

effets de race. La variance normalisée minimale possible pour l'estimation de l'effet du revenu (objectif 3) est de 4 (27 % de la variance sous EAS), valeur obtenue quand le revenu est connu au niveau individuel sous un plan d'échantillonnage dans lequel l'échantillon est réparti de manière égale entre les pauvres et les non-pauvres. Dans le cas d'information au niveau de l'lot, la variance peut être réduite à 74 % de sa valeur sous EAS. Alors que ne connaît que la race offre peu d'avantages pour cet objectif, l'ajout de données individuelles sur la race à l'information au niveau de l'lot donne une réduction supplémentaire de la variance de l'effet estimé du revenu, laquelle atteint 60 % de sa valeur sous EAS. Les variances des estimations de la moyenne et des

**Tableau 1**  
Distributions des races et de la pauvreté chez les personnes de 65 ans et plus, selon le groupe d'îlots de recensement dans le Nord de la Colombie (CV = coefficient de variation)

Blancs Noirs Asiatiques Hispaniques Autre Total				
Pourcentage de la population	65,70	6,40	16,80	8,20
Pourcentage de la population pauvre	5,20	14,20	10,10	10,60
CV pourcentage	0,46	2,94	1,21	1,73
CV pourcentage par îlot de la race	1,53	1,37	1,58	1,89
par îlot de pauvres	2,30	2,28	-	2,30

**3.2 Contraintes du plan d'échantillonnage :**  
information disponible et objectifs du plan

Nous avons calculé l'efficacité, comparativement à l'échantillonnage aléatoire simple (EAS), du plan optimal pour des scénarios définis par deux contraintes, à savoir (1) le choix de la fonction objectif et (2) les hypothèses au sujet de l'information disponible pour déterminer les probabilités d'échantillonnage.

Nous avons considéré six hypothèses possibles au sujet de l'information disponible pour la race (non disponible, ou disponible au niveau individuel) et pour le revenu (non disponible, disponible uniquement par îlot, ou disponible au niveau individuel). Comme la race est plus souvent consignée dans les dossiers d'hospitalisation que le revenu, nous avons exclu le cas où la catégorie de revenu est connue au niveau individuel, mais la race n'est connue que par groupe d'îlots. Chaque hypothèse correspond à une définition de la cellule utilisée pour l'élaboration de la section 2.3 et à une définition correspondante des variables **t** et **u** :

1. Aucune information sur les variables du plan d'échantillonnage n'est disponible : la cellule correspond à l'ensemble de la population et **u** inclut la race et le revenu. (Colonnes intitulées « EAS » dans le tableau 2.)
2. Race uniquement : la cellule est une catégorie raciale, **u** contient les variables de race, et **t** est le revenu. (Colonnes intitulées « Race ».)
3. Données agrégées pour l'îlot uniquement : la cellule est un groupe d'îlots de recensement, **u** est vide et **t** comprend la race et le revenu. (Colonnes intitulées « îlot ».)
4. Données individuelles sur la race, données agrégées au niveau de l'îlot sur le revenu selon la race : la cellule correspond à la population d'une race dans un groupe d'îlots, **u** est la race et **t** est le revenu. (Colonnes intitulées « Race+îlot ».)
5. Données individuelles sur le revenu, aucune donnée sur la race : la cellule est une catégorie de revenu, **u** est le revenu et **t** est vide. (Colonnes intitulées « Revenu ».)

6. Données sur la race et sur le revenu disponibles pour chaque individu : la cellule est une catégorie raciale selon le revenu, **u** comprend la race et le revenu et **t** est vide. (Colonnes intitulées « Race+Revenu ».)

Nous avons calculé les taux d'échantillonnage optimaux sous chaque hypothèse au sujet de l'information disponible, en utilisant diverses fonctions objectif. Dans chacune de ces fonctions, nous avons pondéré ensemble les variances des estimations des coefficients dans certains de quatre modèles de régression ou tous : (1) le modèle à « terme constant uniquement » dont le seul paramètre est la moyenne de la population, (2) un modèle de race paramétrisé comme une

moyenne de la population, (3) un modèle de revenu paramétrisé comme une moyenne de la population, et (4) un modèle additif conjoint comprenant des effets de race et de revenu. Chaque fonction objectif comprend la pondération  $d^{moyenne}_{ij}$ , qui garantit que toutes les probabilités  $\pi_{ij} > 0$ , ce qui évite des problèmes numériques dans l'optimisation. Donc, au moins deux modèles sont représentés dans chaque fonction objectif (section 2.4). Quand la fonction objectif pondère à la fois les effets de revenu et de race, l'effet de revenu unique reçoit la pondération  $d^{revenu}_{ij} = 3$  pour qu'elle concorde avec les trois effets de race ayant une pondération de 1.

Nous avons étudié un choix de poids objectifs qui mettent l'accent sur l'estimation des effets de race, des effets de revenu ou des deux. Chaque partie du tableau 2 représente un choix particulier de poids objectifs  $d_i$  (troisième colonne) pour les coefficients des contrastes  $a_i$  (deuxième colonne) d'une série de modèles (premières colonnes). La quatrième colonne donne la variance (normale) pour une taille d'échantillon unitaire  $n/N_{ij}$  pour l'estimation de ce coefficient sous EAS en supposant que la variance résiduelle est  $\sigma^2 = 1$ . Dans les autres colonnes, nous présentons les effets de plan, les ratios de la variance normalisée  $n/N_{ij}$  pour le plan optimisé sous diverses hypothèses au sujet de l'information disponible sur les variables du plan par rapport à la variance sous EAS. Les lignes dont le poids objectif est  $d_i = 0$  n'ont pas d'incidence sur l'optimisation, mais sont incluses pour illustrer l'effet de chaque plan d'échantillonnage sur l'efficacité de l'estimation d'un coefficient qui n'est pas inclus dans la fonction objectif. La dernière ligne résume l'effet de plan pondéré correspondant à la fonction de perte, c'est-à-dire la combinaison pondérée des variances.



3. Application : Régressions sur la race et l'état de pauvreté

3.1 Description des emplacements et des données

Le projet CanCORS (Ayanian et coll. 2004) comprend cinq emplacements définis géographiquement (Nord de la Californie, Los Angeles, Alabama/Géorgie, Caroline du Nord et Iowa) et deux emplacements définis sur une base organisationnelle. Le Nord de la Californie comprend neuf comtés qui s'étendent de la région de la baie de San Francisco au comté semi-rural Placer County sur la frontière avec le Nevada. Cet emplacement est diversifié ethniquement et géographiquement et, par conséquent, convient bien pour illustrer les méthodes. Nous décrivons les résultats pour cet emplacement en détail, puis nous résumons les résultats pour les autres.

Nos données sont fondées sur l'échantillon qui a reçu le « questionnaire détaillé » du Recensement décennal des États-Unis de 2000 et ont été extraites pour les neuf comtés de notre région cible (Alameda, Contra Costa, Placer, Sacramento, San Francisco, San Joaquin, San Mateo, Santa Clara et Solano) de SF-3, tableaux 159a à 159i, « Poverty Status in 1999 by Age ». Nous avons produit des tableaux croisés pour les résidents d'au moins 65 ans échantillonnés de chaque groupe d'îlots de recensement (une petite région de chaque îlot de recensement à plusieurs îlots urbains, appelée ci-après un îlot) selon la race/ethnicité et le revenu, en utilisant les poids d'échantillonnage du recensement. La contrainte d'âge correspond approximativement aux âges pour lesquels sont relevés le plus de nouveaux cas de cancer

admissibles pour l'étude. Le revenu du ménage a été dichotomisé comme étant supérieur ou inférieur au seuil de pauvreté standard. Le recensement comportait des questions distinctes sur l'ethnicité hispanique et la race; nous avons classifié la population comme étant hispanique ou blanche non hispanique, noire, ou asiatico-américaine. Une catégorie hétérogène « Autre » regroupe les 3 % restant de la population de personnes âgées. (Par souci de concision, dans la suite de l'exposé, nous parlerons de « catégorie raciale ».) La région étudiée comptait 844 560 personnes de plus de 65 ans réparties dans 5 098 groupes d'îlots, soit une moyenne de 166 par groupe d'îlots.

Le tableau 1 résume la répartition selon la race et le revenu dans le Nord de la Californie. Les Noirs sont les individus dont le taux global de pauvreté est le plus élevé et chez lesquels la ségrégation est la plus forte (coefficient de variation le plus élevé du pourcentage de Noirs par îlot), résultats qui confirment les profils nationaux de ségrégation résidentielle (Denton et Massey 1993). Les Hispaniques sont ceux pour lesquels la variation géographique relative du taux de pauvreté est la plus prononcée (coefficient de variation le plus élevé du taux de pauvreté par îlot).

autre, dans le même modèle ou dans des modèles différents. Faire de la moyenne de population l'une des grandeurs à estimer facilite le contrôle de ce risque. Nous illustrons ce point à l'aide d'un exemple où les variables de plan d'échantillonnage  $x_k, z_k$  qui suivent une loi conjointe

$$\zeta : (X, Z) \sim N(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}),$$

sont entièrement observées pour les individus (dénotés par l'indice  $k$  à la section 2.2) et les modèles de régression constante et de régression univariée :

$$\text{Modèle 0 : } y_k = \alpha_0 + \varepsilon_k^0 \quad \varepsilon_k^0 \sim [0, \sigma_0^2]$$

$$\text{Modèle 1 : } y_k = \alpha_x + \beta_x x_k + \varepsilon_k^x \quad \varepsilon_k^x \sim [0, \sigma_x^2]$$

Modèle 2 :  $y_k = \alpha_z + \beta_z z_k + \varepsilon_k^z, \varepsilon_k^z \sim [0, \sigma_z^2]$ . Pour simplifier la notation, nous supposons que  $\sigma_0^2 \approx \sigma_x^2 \approx \sigma_z^2 \approx 1, x_U = z_U = 0$  et que les coûts  $c_k$  sont constants.

Considérons le plan d'échantillonnage optimisé pour  $V =$

$dV(\hat{\alpha}_0) + V(\hat{\beta}_x), d \geq 0$ . En vertu de (4), le plan optimal a pour probabilités  $\pi_k \propto \sqrt{d + x_k^2}$ . Sous ce plan, la variance prévue est approximée par  $V(\hat{\beta}_z) \approx n^{-1} \sigma_z^2 (Z_U^T Z_U)^{-1} = (Z_U^T W_U Z_U) (Z_U^T Z_U)^{-1}$  où  $Z_U = (z_1^N, \dots, z_N^N)^T$  et  $W_U = \text{diag}(\pi_1^2, \dots, \pi_N^2)$ . Alors,  $E_{\pi} n V(\hat{\beta}_z) \approx c_0 E_{\pi} (Z_U^T / \sqrt{d + X_U^2}) (Z_U^T / \sqrt{d + X_U^2})$  où  $c_0$  dépend uniquement de  $d$ , de sorte que  $E_{\pi} n V(\hat{\beta}_z)$  dépend de  $p$  et  $d$ . Si  $d = 0$  (aucun poids n'est lié à l'estimation de la moyenne),  $E_{\pi} n V(\hat{\beta}_z) = \infty$ , à moins que  $p = \pm 1$ . Donc, sauf si l'objectif accorde un certain poids à la variance de l'estimateur de la moyenne, le plan pourrait être très médiocre pour les coefficients liés aux covariables qui ne se trouvent pas dans l'étendue des variables des modèles optimisés. Par contre, si  $d > 0$ , nous pouvons décomposer  $Z$  en composantes parallèles et perpendiculaires à  $X, Z = Z_1 + Z_2$ , où  $Z_1 = pX$  et  $Z_2 = Z - pX$ , de sorte que  $Z_1 \perp Z_2, Z_1 \perp X$  et  $E_{\pi} Z_2 = 0$ . Alors,

$$E_{\pi} n V(\hat{\beta}_z) = p^2 c_0 E_{\pi} (X_U^T / \sqrt{d + X_U^2}) (X_U^T / \sqrt{d + X_U^2}) + (1 - p^2) c_0 E_{\pi} (1 / \sqrt{d + X_U^2}) = p^2 E_{\pi} Z_1^T E_{\pi}^{\text{opt}} Z_1 (1 - p^2) E_{\pi}^{\text{opt}} (1 / \sqrt{d + X_U^2}).$$

Exprimé en mots, la variance du coefficient du nouveau modèle est une combinaison des deux variances qui étaient contrôlées dans l'optimisation. Cela donne à penser qu'un plan d'échantillonnage qui comprend l'estimation de la moyenne globale de population dans l'optimisation offre une certaine protection contre l'inefficacité extrême pour d'autres modèles contenant des variables qui n'ont pas été prises en compte dans le plan original, quoique les résultats simples présentés ici ne se généralisent pas nécessairement de manière à couvrir chaque cas.



données sont disponibles uniquement pour un échantillon, (2) les données sur le plan d'échantillonnage sont anciennes et la distribution des variables du plan d'échantillonnage dans la cellule pourrait avoir changé au cours du temps ou (3) les données sur les éléments individuels sont mesurées avec une erreur. De même, la distribution de  $t$  pourrait n'être disponible que pour une supercellule qui contient plusieurs valeurs de  $\mathbf{u}$  (par exemple, la race et l'îlot de recensement de résidence sont connus pour chaque individu, mais la distribution du revenu est connue pour l'îlot dans son ensemble, mais non pour chaque race dans l'îlot), si bien que  $t_0$  et  $S_{T_0}$  doivent être estimés sous un modèle.

2.4 Plus d'un modèle

Dans l'exposé qui précède, nous avons supposé que toutes les grandeurs que l'on veut estimer sont des combinaisons des paramètres d'un seul modèle. De manière plus générale, les analyses envisagées pourraient comprendre l'ajustement de plusieurs modèles, et  $V$  pourrait représenter la somme des variances de diverses combinaisons de paramètres provenant de ces modèles. Un cas particulier évident est l'estimation d'une moyenne de population (comme il est suggéré à la section 2.2), le coefficient du modèle  $y_k = \beta_0 \cdot 1 + \varepsilon_{k0}$ , ainsi que certains coefficients de régression. Un autre exemple simple est l'estimation des moyennes de divers domaines définis, c'est-à-dire les coefficients de modèles de la forme  $y_k = \beta_0^{(m)} x^{(m)bk} + \varepsilon_{k0}^{(m)bk}$ , où  $x^{(m)bk}$  est un vecteur d'indicateurs d'appartenance à un domaine, les diverses définitions de domaine étant désignées par l'indice  $m = 1, \dots, M$ , ou les contrastes de ces moyennes. Par exemple, nous pourrions souhaiter estimer les moyennes selon la race ainsi que selon l'âge.

Si chacune des combinaisons d'intérêt ne comprend que des paramètres en provenance d'un seul modèle, alors chaque combinaison possède sa propre matrice de plan, de sorte que l'indice de modèle  $m$  peut être identifié à l'indice de la grandeur à estimer  $i$ . Donc, dans (5) et (6), nous remplaçons  $X_U$  par  $X_U^{(i)}$ , ainsi que  $S_b$  par  $S_b^{(i)}$ .

Si certaines grandeurs à estimer combinent des paramètres provenant de modèles différents, nous employons les estimateurs  $\beta^{(m)}$  pour les divers modèles. Puis, dans (5) et (6), nous remplaçons  $X_U' X_U$  par  $\text{diag}(X_U'^{(1)} X_U^{(1)}, \dots, X_U'^{(M)} X_U^{(M)})$ ,  $S_b$  comme étant les sommes combinées de la matrice des carrés et des produits croisés pour l'ensemble des modèles, avec les blocs

$$S^{b(m), m'} = \begin{pmatrix} \mathbf{u}^{b(m)} \mathbf{u}^{b(m')} \\ \mathbf{u}^{b(m)} \mathbf{u}^{Tb(m')} \end{pmatrix} S^{Tb(m), m'}$$

La suite de l'optimisation est la même qu'à la section 2.3.

2.5 Poids de contraste flexibles

Dans l'étude CanCORS, nous voulons estimer l'effet du revenu en tenant compte de l'effet de la race et de l'effet moyen pour l'ensemble des races. La façon dont les races sont pondérées dans cette moyenne est moins importante pour nous, car les régions étudiées ne sont pas représentatives des proportions nationales selon la race. Nous pourrions alors estimer l'effet du revenu sur la pauvrete/non-pauvrete pour chaque race et combiner les résultats obtenus en utilisant des poids choisis de façon à minimiser la variance de l'estimateur de la moyenne pondérée des effets intra-race du revenu.

En général, nous examinons des situations dans lesquelles l'intérêt scientifique est d'estimer ou de tester toute combinaison  $\mathbf{a}_j = \mathbf{A}_j \mathbf{f}_j$ , où  $\mathbf{A}_j$  est fixe et tous les  $\mathbf{f}_j$  sont arbitraires (et ne sont pas nécessairement tous de la même dimension) sous la contrainte  $\|\mathbf{f}_j\|_1 = 1$ ,  $f_{jg} \geq 0$ . Dans notre exemple motivé par l'étude CanCORS, le modèle sous-jacent comprend huit variables indicatrices pour chacun des groupes définis par quatre groupes raciaux croisés avec un niveau de pauvreté dichotomique, et  $\mathbf{A}_j$  est une matrice  $8 \times 4$  dont chaque colonne contient une valeur 1 et une valeur -1 pour le contraste entre pauvre et non pauvre au sein d'une race. Alors,  $\mathbf{f}_j$  contient les poids appliqués au contraste dans chaque race, et  $\mathbf{a}_j = (f_{j1}, \dots, f_{j11}, -f_{j12}, \dots, -f_{j13}, -f_{j14}, \dots, -f_{j14})$  est le contraste pondéré des coefficients des huit indicateurs.

En substituant dans (5) et (6), nous optimisons sur les probabilités d'échantillonnage  $\pi = \{\pi_k\}$  ainsi que les poids de combinaison  $\mathbf{f} = \{\mathbf{f}_j\}$ . Dans le cas de modèles multiples, nous utilisons l'une ou l'autre des formules de la section 2.4, selon que les combinaisons d'intérêt comprennent les coefficients d'un ou de plusieurs modèles. La définition de  $\mathbf{a}_j$  est donc déterminée en partie par des considérations scientifiques et en partie par l'information disponible provenant de la population étudiée.

Une approche naturelle de l'optimisation conjointe de  $\pi$  et  $\mathbf{f}$  consiste à minimiser alternativement  $V$  par rapport à  $\pi$  en utilisant l'expression (6) modifiée et par rapport à  $\mathbf{f}$ , en observant les contraintes sur  $\mathbf{f}$ . Dans l'optimisation,  $\mathbf{f}_j$  apparaît dans une expression de la forme  $\mathbf{f}_j' \mathbf{D}_j(\pi) \mathbf{f}_j$ . Par minimisation sous la contrainte  $\mathbf{f}_j' \cdot \mathbf{1} = 1$  en utilisant les multiplicateurs de Lagrange, nous obtenons  $\mathbf{f}_j = \mathbf{D}_j'(\pi) \mathbf{1} / (\mathbf{1}' \mathbf{D}_j'(\pi) \mathbf{1})$  à condition que  $\pi_{k0} > 0$  et que les contraintes de non-négativité ne soient pas obligatoires. Si elles le sont, des méthodes de programmation quadratique peuvent être utilisées.

2.6 Précision des analyses non prévues

Un plan d'échantillonnage destiné à être optimal pour un coefficient de régression pourrait être très inefficace pour un

Dans le contexte de l'inférence fondée sur le plan de sondage, la mesure pertinente est la variance moyenne sous le plan d'échantillonnage sur toutes les populations possibles obtenues sous le modèle  $E[\text{Var}_\pi(\mathbf{a}^T \boldsymbol{\beta})]$  (« variance attendue » d'Isaki et Fuller 1982; voir aussi Bellhouse 1984, section 1); cette grandeur est approximativement égale à la variance sous un modèle prévu (voir l'annexe pour la preuve et les conditions asymptotiques).

En vertu de l'argument type du multiplicateur de Lagrange pour les problèmes de répartition optimale (par exemple, Valliant, Dorfman et Royall 2000, pages 169-

170),  $V_\pi$  est minimisée sachant la contrainte de coût attendu  $(1) \text{ quand } \partial V_\pi / \partial \pi^k = c^k \lambda$  pour une constante  $\lambda$ , et toutes les valeurs de  $k$ , de sorte que  $\pi^k \propto \sigma^k (\mathbf{a}^T \mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{X}_k^k / \sqrt{c^k}$ . Donc, le taux d'échantillonnage optimal est plus élevé pour les cas où la variance sous le modèle est plus grande et le coût est plus faible (comme dans le cas habituel d'estimation d'une moyenne), ainsi que pour les cas dont l'effet levier est plus grand dans la régression. Ce résultat diffère des calculs fondés sur un modèle standard pour obtenir un plan expérimental optimal, qui donnerait la répartition de l'échantillon complet entre quelques points du plan à effet levier élevé. L'estimateur convergent sous le plan de la régression GDP ne s'appuie pas sur l'hypothèse que le modèle est correct et, par conséquent, requiert que chaque cas ait une probabilité de sélection positive. Donc, pour l'estimation d'un ratio  $\beta$  sous un modèle homoscedastique  $y_k = \beta x_k + \varepsilon_k$ , l'estimation fondée sur un modèle semblerait dicter la sélection des unités ayant les plus grande valeur de  $x_k$ , mais nos probabilités de sélection sont proportionnelles à  $x_k$ .

Habituellement, une étude a pour but d'estimer plus d'une grandeur; ainsi, l'étude CamCORS est conçue pour estimer les effets de la race ainsi que du revenu. Nous généralisons (3) à l'estimation simultanée de plusieurs combinaisons linéaires de coefficients en optimisant une somme pondérée des variances  $V = \sum_i d_i' V_{\pi^i}^{-1} d_i$  où l'indice  $i$  désigne les grandeurs à estimer. Suivant les mêmes arguments, les probabilités d'échantillonnage optimales pour cet objectif sont

$$\pi^k \propto \sigma^k \left( \sum_i d_i' (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{X}_k^k \right)^2 / c_k^{1/2} \quad (4)$$

Pour certains choix de  $\{\mathbf{a}_i\}$ , l'adhésion stricte à l'expression (4) pourrait produire des probabilités  $\pi^{k_k}$  arbitrairement faibles (et, donc, des poids arbitrairement grands) pour les cas dont l'effet levier est presque nul. Afin d'éviter cette situation, nous pourrions imposer une valeur plancher positive à  $\pi^k$ . Ou bien, en incluant l'estimation de la moyenne de population parmi les objectifs (section 2.4), nous garantissons un échantillonnage avec probabilité

$$\pi^k \propto \sigma^k \left( \sum_i d_i' \mathbf{a}_i' (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{S}_b (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{a}_i / K c_k^{1/2} \right) \quad (6)$$

est la matrice des carrés moyens et des produits croisés dans la cellule  $b$ , avec  $\mathbf{S}^{Tb} = \mathbf{T}^{Tb} / K$ . Les probabilités optimales d'échantillonnage correspondant à (4) sont alors

$$\mathbf{S}^b = \begin{pmatrix} \mathbf{u}_1^b & \mathbf{u}_2^b & \mathbf{u}_3^b \\ \mathbf{z}_1^b & \mathbf{z}_2^b & \mathbf{z}_3^b \end{pmatrix} \mathbf{S}^{Tb}$$

Si  $\mathbf{t}$  est mesuré par un recensement dans chaque cellule, alors  $\mathbf{t}_b$  et  $\mathbf{S}^{Tb}$  sont connus exactement. Les mêmes principes s'appliquent toutefois si, au lieu d'être observé directement,  $\mathbf{S}^{Tb}$  est estimé sous un modèle  $\zeta$ . Nous remplaçons alors  $\mathbf{t}_b$  et  $\mathbf{S}^{Tb}$  dans (5) par les espérances prédictives  $\tilde{\mathbf{t}}_b = E_{\zeta} \mathbf{t}_b$  et  $\tilde{\mathbf{S}}^b = E_{\zeta} \mathbf{S}^{Tb}$ . Les situations qui suivent pourraient être des exemples : 1) pour chaque cellule, les

### 2.3 Plan optimal avec variables aux niveaux individuel et agrégé

Supposons maintenant que le vecteur de covariables  $\mathbf{t}^{bk}$  n'est pas vide et que  $\mathbf{u}^{bk}$  est constant dans chaque cellule, comme quand l'information agrégée sur le plan est disponible pour les cellules correspondant à des classes de covariables de  $\mathbf{u}$  à l'intérieur des îlots. Dans l'étude CamCORS, si nous connaissons la race ( $\mathbf{u}$ ), mais non le revenu ( $\mathbf{t}$ ) pour les sujets pris individuellement et que nous connaissons la distribution du revenu pour chaque race dans chaque îlot de recensement, nous définirons des cellules constituées de personnes d'une seule race dans un seul îlot de recensement.

Puisqu'il est impossible de faire la distinction entre les cas appartenant à une même cellule d'après les covariables, nous supposons en outre que  $\sigma^{bk} = \sigma^b$  et  $c^{bk} = c^b$  sont constants dans toute la cellule, de sorte que le plan optimal rend aussi  $\pi^{bk} = \pi^b$  constante dans chaque cellule. Nous pouvons maintenant réécrire (3) sous la forme

$$V_\pi \approx \mathbf{a}^T (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{E}[\text{Var}_\pi^S \sum_k x_k^{bk} y_k^k / \pi^b] (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{a} = \mathbf{a}^T (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \left( \sum_k (\sigma_y^k / \pi^b) \mathbf{S}_b^k \right) (\mathbf{X}_U^U \mathbf{X}_U^U)^{-1} \mathbf{a} \quad (5)$$

où

sondage dans la base de sondage, puis nous généralisons à des situations où l'information sur certaines variables n'est disponible que sous forme d'agrégat pour des sous-domaines ou provient de données inexactes. Ensuite, nous envisageons l'estimation optimale de combinaisons de coefficients provenant de plus d'un modèle et de combinaisons flexibles de coefficients. À la section 3, nous estimons les avantages possibles de ces méthodes pour un sondage réalisé aux emplacements choisis pour l'étude CanCORS, en utilisant des données de recensement au niveau de l'ilot sur la race/ethnicité et sur la pauvreté. Enfin, à la section 4, nous examinons la pertinence de l'approche GDP et les extensions possibles de la méthodologie.

## 2. Calcul du plan de sondage optimal

### 2.1 Notation

Supposons que la population cible est divisée en cellules dénotées par les indices  $b = 1, 2, \dots, B$ , avec les éléments de la cellule  $b$  dénotés par les indices  $k = 1, 2, \dots, K_b$ . À chaque élément est associé un vecteur de covariables  $\mathbf{x}^{bk}$  avec  $\mathbf{x}^{bk} = (\mathbf{u}^{bk}, \mathbf{t}^{bk})$ , où  $\mathbf{u}^{bk}$  est la composante observée pour les individus identifiables. La distribution de  $\mathbf{t}^{bk}$  est connue dans chaque cellule, mais les valeurs individuelles ne sont pas observées; donc, la cellule est l'unité d'agrégation pour certaines ou pour toutes les variables du plan. Par conséquent, nous connaissons les valeurs de population finie  $\mathbf{T}_b = (\mathbf{t}_b^1, \mathbf{t}_b^2, \dots, \mathbf{t}_b^{K_b})$ , mais nous ne pouvons pas associer les lignes avec les individus. Définissons  $\mathbf{t}_b = \mathbf{T}_b / K_b$ , la moyenne de  $\mathbf{t}$  dans la cellule  $b$ .

Un coût  $c^{bk}$  est associé à l'échantillonnage de chaque élément. Un plan d'échantillonnage est défini en attribuant une probabilité de sélection  $\pi^{bk}$  à chaque élément. Supposons une contrainte sur le coût attendu,

$$(1) \quad \sum_{b,k} c^{bk} \pi^{bk} \leq C.$$

Afin de simplifier l'exposé, nous supposons aussi que le taux d'échantillonnage est faible et que les avantages éventuels de la stratification sont minimes, si bien que le plan peut être décrit approximativement comme un échantillonnage avec probabilités inégales non stratifié avec remise. Nous supposons également qu'il s'agit d'un échantillonnage par élément à un seul degré. La population est  $U = \{b, k : b = 1, 2, \dots, B; k = 1, 2, \dots, K_b\}$  et un échantillon est  $S \subset U$ . Le coefficient de régression par les moindres carrés ordinaires (MCO) descriptif de la population, qui correspond au modèle  $y^{bk} = \beta' \mathbf{x}^{bk} + \varepsilon^{bk}$  avec  $\varepsilon^{bk} \sim [0, \sigma^2]$ , est  $\beta^U = (\mathbf{X}'^U \mathbf{X}^U)^{-1} \mathbf{X}'^U \mathbf{y}^U$ , où l'indice inférieur  $U$  indique que les matrices ou les vecteurs correspondent à l'ensemble de la

population. Ici  $[0, \sigma^2]$  signifie une loi de moyenne 0 et de variance  $\sigma^2$ , mais de forme non spécifiée.) Alors

$$(2) \quad \hat{\beta} = (\mathbf{X}_S' \mathbf{W}_S \mathbf{X}_S)^{-1} \mathbf{X}_S' \mathbf{W}_S \mathbf{y}_S$$

est l'estimateur sous le plan de sondage habituel de  $\beta$ , où  $S$  signifie que seules les lignes correspondant à l'échantillon sont incluses, et  $\mathbf{W}$  est la matrice diagonale des poids  $1/\pi^{bk}$ .

Pour concevoir le sondage, nous devons émettre certaines hypothèses quant à la distribution des résultats  $y^{bk}$ , même si nous ne nous appuyons pas sur les mêmes hypothèses pour analyser les données. Plus précisément, nous supposons que les résultats sont générés par un modèle indépendant dont les  $\varepsilon^{bk} \sim [0, \sigma^2]$  sont indépendantes et les  $\sigma^{bk}$  sont connues (jusqu'à un facteur constant). Notons que les distributions des variables du plan d'échantillonnage  $\mathbf{x}^{bk}$  et des résidus importent pour l'optimisation du plan de sondage, mais non la valeur de  $\beta$ , parce qu'elle n'a pas d'incidence sur la variance des estimateurs par la régression. En outre, l'hypothèse d'indépendance des résidus d'un modèle de régression pourrait être plus raisonnable que celle de l'homoscédasticité, même lorsque nous tenons compte de l'homoscédasticité, même lorsque nous ajustons un modèle MCO. Les coefficients obtenus par les MCO (y compris les cas particuliers tels que la moyenne globale ou les moyennes de domaine) sont souvent des statistiques descriptives utiles, même si le modèle MCO ne tient effectivement pas; toutefois, si l'information au sujet de l'hétéroscédasticité est disponible, elle peut être utilisée pour rendre le plan de sondage plus efficace.

### 2.2 Plan de régression GDP optimal avec variables au niveau individuel uniquement

Considérons d'abord le cas où  $\mathbf{t}$  est vide, de sorte que  $\mathbf{x}^{bk} = \mathbf{u}^{bk}$ , traduisant un scénario où le chercheur possède l'information sur toutes les variables du plan d'échantillonnage pertinentes (race et revenu dans notre plan CanCORS) avant l'échantillonnage. Puisque, dans ces conditions, les cellules correspondent à des cas uniques, nous laissons tomber l'indice inférieur  $b$  et écrivons  $\hat{\beta} = (\sum_S \mathbf{w}_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_S \mathbf{w}_k \mathbf{x}_k y_k$ . Alors, pour toute combinaison linéaire fixe de coefficients avec les poids  $\mathbf{a}$ , en supposant que le premier facteur est un estimateur convergent sous le plan (après mise à l'échelle) de  $N(\mathbf{X}'^U \mathbf{X}^U)^{-1}$ , nous obtenons l'espérance sous échantillonnage de la variance fondée sur le modèle (White 1980) de l'estimateur.

$$\begin{aligned} V^a &= E^u \text{Var}^u \hat{\beta} \\ &\approx \mathbf{a}' (\mathbf{X}'^U \mathbf{X}^U)^{-1} E^u \text{Var}^u \sum_{k \in S} \mathbf{x}_k y_k / \pi_k (\mathbf{X}'^U \mathbf{X}^U)^{-1} \mathbf{a} \\ &= \mathbf{a}' (\mathbf{X}'^U \mathbf{X}^U)^{-1} \left( \sum_{k \in U} (\sigma_k^2 / \pi_k) \mathbf{x}_k \mathbf{x}_k' \right) (\mathbf{X}'^U \mathbf{X}^U)^{-1} \mathbf{a}. \end{aligned} \quad (3)$$



population; dans le cas de la régression, nous parlerons de régression d'une « grandeur descriptive de la population » (*descriptive population quantity*) ou GDP (Pfeffermann 1993, pages 319 à 321).

Depuis Neyman (1934) l'abondante littérature sur la conception optimale des sondages (recensée dans des ouvrages classiques tels que Cochran 1977 ou Särndal, Swensson et Wretman 1992) a porté principalement sur l'estimation d'une grandeur simple, telle qu'une moyenne ou un ratio, ou de plusieurs grandeurs de ce genre (Kish 1974; Bellhouse 1984; Chromy 1987). Bien que l'estimation de la variance des estimations fondées sur le plan de sondage des coefficients de régression ait fait couler beaucoup d'encre (Fuller 1975; Fuller 1984; Binder 1981; Binder 1983), assez peu d'attention a été accordée aux plans d'échantillonnage optimaux correspondants. L'estimation assistée par la régression d'une moyenne (Cassel, Särndal et Wretman 1976; Särndal, Swensson et Wretman 1992, section 12.2) est un problème distinct.]

De surcroît, les caractéristiques qui pourraient être utilisées pour définir un plan d'échantillonnage à probabilités inégales sont vraisemblablement enregistrées avec une erreur dans les bases de sondage, parce qu'elles sont fondées sur des données agrégées ou que les caractéristiques associées à une unité (comme une adresse ou un ménage) changent au cours du temps. Ce genre d'erreur peut affecter fortement l'efficacité d'un plan d'échantillonnage qui est censé être optimal, voir Morris, Newhouse et Archibald (1979, section III) au sujet de l'échantillonnage stratifié pour l'estimation par domaine et Thomsen, Tesfu et Binder (1986) au sujet de l'échantillonnage avec probabilité proportionnelle à la taille. Waksberg (1973, 1995) considère la stratification par îlot de recensement sur une seule caractéristique agrégée pour l'estimation des moyennes de domaines tels que des groupes raciaux/ethniques ou les pauvres, avec ou sans intervention de sélection subséquente.

Notre objectif ici est de décrire des plans optimaux pour des échantillons qui seront utilisés en analyse par régression (pondérée par les poids de sondage) d'une grandeur descriptive de la population (GDP), au sens de la minimisation de la somme pondérée des variances de certaines combinaisons linéaires présélectionnées de coefficients de régression. Nous examinons aussi certaines classes de grandeurs à estimer et les estimateurs correspondants qui s'écartent de l'approche GDP afin d'accroître l'efficacité. À la section 2, nous établissons la notation et dérivons les laux d'échantillonnage optimaux pour la régression GDP sous des scénarios représentatifs de l'information au niveau individuel et au niveau du domaine dont on pourrait disposer dans des sondages de population quand la base de sondage est imparfaite. Nous supposons d'abord que nous disposons d'information exacte sur les variables du plan de

être combinées à des données de recensement sur les distributions race-revenu dans les îlots de recensement pour procéder à un échantillonnage non proportionnel des patients et, donc, améliorer les estimations des effets de la

race et du revenu. Ce genre de préoccupations surviennent souvent quand des données de sondage sont utilisées pour estimer les coefficients d'un ou de plusieurs modèles de régression. Par exemple, la National Health Interview Survey (NHIS) comporte un suréchantillonnage géographique conjugué à une interview de sélection pour suréchantillonner les répondants noirs et hispaniques de manière à améliorer l'estimation par domaine (Botman, Moore, Moriarty et Parsons 2000, page 12); les données de la NHIS ont été utilisées à grande échelle dans des analyses par régression, dont l'estimation par domaine est un cas particulier. Sasthy, Ghosh-Dastidar, Adams et Pebley (2005, pages 1013-1014) ont suréchantillonné des secteurs de recensement en fonction de leur composition en groupes minoritaires, en recourant à des simulations pour évaluer la puissance de divers plans d'échantillonnage pour les analyses par régression d'intérêt. Le Youth Risk Behavior Surveillance System suréchantillonne les *écoles* dans les UPE où la prévalence des minorités est forte afin d'améliorer la précision des estimations pour les groupes raciaux/ethniques minoritaires (Eaton, Kann, Kinchen, Ross, Hawkins, Harris, Lowry, McManus, Chyen, Shanklin, Lim, Grunbaum et Wechsler 2006, pages 2-3).

La littérature sur la conception optimale des expériences (Royall 1970), mais l'application de ces principes à la conception des sondages est controversée (voir la section 4). L'approche fondée sur le plan de sondage requiert un échantillon qui est représentatif, grâce à un mécanisme probabiliste connu, d'une population définie; des positions intermédiaires sont également possibles (sections 2.5, 3-4). De ce point de vue, l'échantillonneur n'est pas libre de sélectionner, par exemple, 100 répondants blancs dans un quartier dont, de façon commode, la population est principalement blanche et 100 répondants hispaniques dans un quartier dont, de façon commode, la population est principalement hispanique et dire que l'échantillon est « représentatif » pour l'estimation des différences entre les Blancs et les Hispaniques. Les objets de l'inférence fondée sur le plan de sondage sont des grandeurs qui décrivent la

# Répartition optimale de l'échantillon pour une régression convergente sous le plan dans une enquête sur les services aux cancéreux quand les variables du plan d'échantillonnage sont connues pour des agrégats

Alan M. Zaslavsky, Hui Zheng et John Adams

## Résumé

Nous considérons les taux d'échantillonnage optimaux dans des plans d'échantillonnage par élément, quand l'analyse linéaire des coefficients de régression provenant d'un ou de plusieurs modèles. Nous commençons par élaborer des méthodes en supposant que des renseignements exacts sur les variables du plan existent dans la base de sondage, puis nous les généralisons à des situations où l'information pour certaines variables du plan n'est disponible que sous forme d'agrégat pour des groupes de sujets éventuels ou provient de données inexacts ou périmées. Nous envisageons également un plan d'échantillonnage pour l'estimation de combinaisons de coefficients provenant de plus d'un modèle. Une généralisation supplémentaire permet d'utiliser des combinaisons flexibles de coefficients choisis pour améliorer l'estimation d'un effet tout en contrôlant un autre. Les applications éventuelles comprennent l'estimation des moyennes pour plusieurs ensembles de domaines chevauchants, ou l'amélioration des estimations pour des sous-populations telles que les races minoritaires par échantillonnage non proportionnel des régions géographiques. Dans le contexte de la conception d'un sondage sur les soins reçus par les cancéreux (l'étude CanCORS) qui a motivé nos travaux, l'information éventuelle sur les variables du plan d'échantillonnage comprendait des données de recensement au niveau de l'ilot sur la race/ethnicité et la pauvreté, ainsi que des données au niveau individuel. Pour un échantillonnage de l'étude, un plan d'échantillonnage avec probabilités inégales en utilisant les adresses résidentielles des sujets et des données de recensement réduirait la variance de l'estimateur d'un effet du revenu de 25 %, ou de 38 % si la race des sujets avait été connue également. Par pondération flexible des contrastes du revenu selon la race, la variance de l'estimateur serait réduite de 26 % en utilisant les adresses résidentielles seulement et de 52 % en utilisant les adresses et les races. Nos méthodes seraient utiles dans les études où l'on considère un suréchantillonnage géographique selon la race-ethnicité ou les caractéristiques socioéconomiques, ou dans toute étude où les caractéristiques pour lesquelles des données sont disponibles dans les bases de sondage sont mesurées avec une erreur.

Mots clés : Grandeur descriptive de population; erreur de mesure; répartition de Neyman; modèles de régression; plan d'échantillonnage; sondage.

## 1. Introduction

Un sondage doit être conçu afin d'obtenir des données qui serviront à estimer les coefficients d'un ou de plusieurs modèles de régression. On dispose d'information sur la distribution des covariables dans la population et la base de sondage contient également certains renseignements sur les covariables. Comment cette information peut-elle être utilisée pour rendre le plan de sondage plus efficace? De combien un tel plan de sondage peut-il réduire la variance comparativement à l'échantillonnage aléatoire simple et comment varierait cette réponse si la base de sondage fournit les distributions des covariables agrégées sur des groupes, mais non pour les sujets pris individuellement?

Ces questions ont été soulevées au moment de la conception d'un sondage sur les processus de fourniture de soins de santé (tels que la fourniture d'une chimiothérapie quand il convient) et les résultats (tels que la qualité de vie après le traitement) auprès d'un grand échantillon de cancéreux à sept emplacements aux États-Unis mis sur pied

dans le cadre de l'étude CanCORS (Cancer Care Outcomes Research and Surveillance) (Ayanian, Chirischilles, Wallace, Fletcher, Fouad, Kiefe, Hartington, Weeks, Kahn, Malin, Lipscomb, Potosky, Provenzale, Sandler, Vannyn et West 2004). L'un des principaux objectifs de l'étude était d'estimer les effets conjugués de la race et du revenu sur ces mesures, en utilisant des modèles de régression incluant ces deux caractéristiques des patients. Toutefois, on ne disposait que de données limitées au moment de l'échantillonnage des patients en vue de leur participation à l'étude. L'expérience acquise donnait à penser que la race et l'adresse résidentielle pourraient être déterminées de manière raisonnablement exacte au moment de la confirmation des cas en vue de leur recrutement éventuel pour l'étude, mais que le revenu ne pourrait pas être déterminé tant que le sujet n'aurait pas été recruté et interviewé, et que les renseignements ne pourraient pas être recueillis de manière commode durant une entrevue de sélection. Nous avons entrepris les travaux de recherche présentés ici afin de déterminer comment les données disponibles sur les patients pourraient

## Bibliographie

- Lo, A. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, 16, 1684-1695.
- Mak, T.K., et Rao, J.N.K. (1993). A new method for estimating finite population quantiles using auxiliary information. *Canadian Journal of Statistics*, 21, 29-38.
- Meeden, G. (1999). A noninformative Bayesian approach for two-stage cluster sampling. *Sanhnyā, Séries A*, 61, 133-144.
- Meeden, G. (2005). A noninformative bayesian approach to domain estimation. *Journal of Statistical Planning and Inference*, 129, 85-92.
- Nelson, D., et Meeden, G. (1998). Using prior information about population quantiles in finite population sampling. *Sanhnyā, Séries A*, 60, 426-445.
- Nelson, D., et Meeden, G. (2006). Noninformative nonparametric finite population quantile estimation. *Journal of Statistical planning and Inference*, 136, 53-67.
- R Development Core Team (2005). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rao, J.N.K., Kovar, J.G. et Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.
- Rubin, D. (1981). The Bayesian bootstrap. *Annals of Statistics*, 9, 130-134.
- Singh, J.A., Borowsky, S.J., Nugent, S., Murdoch, M., Zhao, Y., Nelson, D., Petzel, R. et Nichol, K.L. (2005). Health related quality of life, functional impairment, and health care utilization by veterans: Veterans' quality of life study. *Journal of American Geriatric Society*, 53, 108-113.
- Smith, R.L. (1984). Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32, 1296-1308.
- Valliant, R., Dorfman, A. et Royall, R. (2000). *Finite Population Sampling and Inference, a Predictive Approach*. New York : John Wiley & Sons, Inc.
- Vardeman, S., et Meeden, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, 12, 675-684.
- Woodruff, R.S. (1952). Confidence intervals for the median and other positive measures. *Journal of the American Statistical Association*, 47, 635-646.
- Zhong, B., et Rao, J.N.K. (2000). Empirical likelihood inference under stratified random sampling using auxiliary population information. *Biometrika*, 87, 920-938.
- Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society, Séries B*, 44, 388-393.
- Booth, J.G., Butler, R.W. et Hall, P. (1994). Bootstrap methods for finite population sampling. *Journal of the American Statistical Association*, 89, 1282-1289.
- Chambers, R.L., et Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.
- Chen, J., et Sitter, S.S. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.
- Chen, J., et Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.
- Cochran, W. (1976). *Sampling Techniques*. New York : John Wiley & Sons, Inc., 3<sup>ème</sup> Edition.
- Deville, J.-C., et Samdal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- Efron, B., et Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, Londres.
- Feller, W. (1968). *An Introduction of Probability Theory and its Applications*. New York : John Wiley & Sons, Inc., Volume I.
- Ghosh, J.K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays*. Par (Ed., D. Basu). New York : Springer-Verlag.
- Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Londres : Chapman and Hall.
- Gross, S. (1980). Median estimation in survey sampling. *Dans Proceedings of the Section of Section on Survey Research Methods*. American Statistical Association, 181-184.
- Hartley, H.O., et Rao, J.N.K. (1968). A new estimation theory for sample surveys. *Biometrika*, 55, 159-167.
- Kazis, L.E., Miller, D.R., Clark, J., Skinner, K., Lee, A., Rogers, W., Spito, A., 3<sup>ème</sup>, Payne, S., Fincke, G., Selim, A. and Linzer, M. (1998). Health related quality of life in patients served by the department of veterans affairs. *Archives of Internal Medicine*, 158, 626-632.
- Kuk, A.Y.C., et Mak, T.K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Séries B*, 51, 261-269.



Pour un échantillon donné  $s = (z_1, \dots, z_k)$ , soit

$$P^s := \{p \mid p \in P \text{ et } p_i = 0 \text{ quand } p_i^s = 0 \text{ pour } i = 1, \dots, k\}. \quad (9)$$

Nous voyons que nous gardons un échantillon si, et

uniquement si,  $P^s$  n'est pas vide.

Dénotons le simplexe à  $k - 1$  dimensions par

$$F := \left\{ (p_1, \dots, p_k) \mid \sum_{i=1}^k p_i = 1, p_i \geq 0 \forall i \in \{1, \dots, k\} \right\}.$$

Pour  $i = 1, \dots, k$ , soit  $e_i$  les sommets de  $F$ . Les  $e_i$  sont les vecteurs unitaires dont la  $i^e$  valeur est 1 et les autres sont 0.

Or,  $P$  est un polytope convexe qui est l'intersection de  $F$  avec l'espace

$$G := \{(p_1, \dots, p_k) \mid A_1 p = b_1, A_2 p \leq b_2\}.$$

Une partition de l'espace des paramètres  $P$  peut être obtenue de la manière suivante. Soit  $F_j$  l'ensemble de

facettes de dimension  $j$  du simplexe  $F$ ,  $j = 0, 1, \dots, k - 1$ . Alors,  $F_0$  est l'ensemble de ses sommets,  $F_j$  est la

série d'enveloppes convexes de toutes les combinaisons de  $j + 1$  sommets, pour  $j = 1, \dots, k - 2$  et  $F_{k-1}$  est le

simplexe  $F$ . Si  $\text{int}(F_j)$  est l'ensemble des intérieurs des facettes de dimension  $j$ , pour  $j = 1, \dots, k - 1$ , alors

$\{F_0, \text{int}(F_1), \dots, \text{int}(F_{k-1})\}$  détermine une partition du simplexe  $F$ . Si  $G_0 := \text{int}(F_0) \cup G$  et  $G_j := \text{int}(F_j) \cup G$

pour  $j = 1, \dots, k - 1$ , alors  $\{G_0, G_1, \dots, G_{k-1}\}$  est une partition de l'espace des paramètres  $P$ . Notons que certains

des  $G_j$  pourraient être vides. Les étapes de l'argument bayésien séquentiel suivent les membres non vides des  $G_j$ .

Si  $Z$  est l'espace échantillon des dénombrements  $(Z_1, \dots, Z_k)$ , alors pour  $p \in P$ , la distribution des dénom-

bremes, disons  $f_p(z) \mid p$  est Multinomiale( $n, p_1, \dots, p_k$ ). quand la taille d'échantillon est  $n$ . Soit  $P_p$  la restriction de l'espace des paramètres  $P$  à  $F_j$ , où ce dernier est un sous-

ensemble de  $P$  et  $Z_{P_p}$  est la restriction de l'espace échantillon  $Z$  déterminée par  $P_p$ .

Nous sommes maintenant prêts à prouver l'admissibilité de l'estimateur fondé sur la loi *a posteriori* de Polya

contre une  $\mu(p)$  sur  $P$ . Supposons que nous sommes à l'étape où nous considérons  $G_j$ , pour un certain  $j \geq 0$ . Supposons que  $G_j = G \cap F$  n'est pas vide pour un

certain sous-ensemble. Deux cas sont possibles.

Cas 1. Si la dimension de  $G_j$  est nulle, c'est-à-dire s'il est

constitué d'un vecteur, disons  $p^0$ , alors nous prenons la loi

*a priori* qui applique une masse unitaire à ce vecteur. Dans

ces conditions, la loi *a posteriori* applique aussi une masse

unitaire à ce vecteur et, si  $z$  est le membre unique de  $Z_{P_p}$

alors l'estimateur bayésien est  $\delta_{p^0}(z) = E(\mu(p) \mid z) = \mu(p^0)$ .

Cas 2. Si la dimension de  $G_j$  est supérieure à zéro, alors la

distribution de  $(Z_1, \dots, Z_k)$  restreinte à  $Z_{P_p}$  est

$$f_{P_p}^p(z \mid p) = \frac{f_p^p(z \mid p)}{f_p^p(z \mid p)} \sum_{z \in Z_{P_p}} f_p^p(z \mid p).$$

La loi *a priori* que nous considérons sur  $P_p$  est

$$\sum_{z \in Z_{P_p}} \pi_{P_p}^p(p) \prod_{i \mid |p| p_i, p_i > 0} d_i,$$

qui peut être normalisée afin d'être une loi *a priori* correcte,

puisque  $\sum_{z \in Z_{P_p}} f_p^p(z \mid p)$  peut s'écrire  $g(p) \prod_{i \mid |p| p_i, p_i > 0} p_i^{p_i}$  où  $g(p)$  est une fonction bornée de  $p$ . Avec cette loi

*a priori*, la loi *a posteriori* est un noyau de la densité de Dirichlet restreint à  $P_p$ .

$$f_p^p(p \mid z) \propto f_p^p(z \mid p) \pi_{P_p}^p(p) \propto \prod_{i \mid |p| p_i, p_i > 0} d_i^{-1}.$$

L'estimateur bayésien de  $\mu(p)$  en fonction de  $\pi_{P_p}^p$ , où  $p$  appartient à  $P_p$ , est  $\delta_{\pi_{P_p}^p}(z) = E(\mu(p) \mid z)$  pour tout  $z$

dans  $Z_{P_p}$ . Donc, si nous utilisons la série de priors

$$\{\pi_{P_p}^p \mid p \in G_0\}, \{\pi_{P_p}^p \mid p \in G_1\}, \dots, \{\pi_{P_p}^p \mid p \in G_{k-1}\},$$

en ignorant à chaque étape les ensembles vides, alors l'estimateur  $\delta(z)$  défini par

$$\delta(z) = \delta_{\pi_{P_p}^p}(z) \text{ pour } z \in Z_{P_p}, F \in G_j, i = 1, \dots, \gamma - 1, (10)$$

où  $\gamma = k$  si  $k < n$  et  $n$  si  $k \geq n$ , est un estimateur admissible de  $\mu(p)$ . Ceci conclut la preuve du théorème

suivant.

**Théorème 1.** Sous le plan d'échantillonnage aléatoire sous contraintes défini par les équations 2 et 3 avec l'espace des paramètres défini par l'équation 8, l'estimateur fondé sur la loi *a posteriori* de Polya contre une population est bayésien séquentiel et donc admissible sous une perte d'erreur quadratique.

## Remerciements

Les travaux de recherche de Glen Meeden et de Radu Lazar ont été financés en partie par la NSF Grant DMS 0406169. Les auteurs remercient les examinateurs et le rédacteur adjoint de plusieurs commentaires constructifs.

comporte exactement comme une loi *a posteriori* bayésienne correcte, mais on ne doit jamais spécifier explicitement de loi *a priori*. Pour plus de détails et la discussion de ces points, consulter Ghosh et Meeden (1997).

Pour prouver l'admissibilité des estimateurs découlant de la loi *a posteriori* de Pôlya pour l'espace des paramètres  $[0, \infty)^N$ , la partie principale de l'argument bayésien séquentiel commence par l'hypothèse que l'espace des paramètres est  $\Delta^N$ , où  $\Delta$  est un ensemble fini arbitraire de nombres réels positifs. Une fois que l'admissibilité a été démontrée pour un espace général  $\Delta$ , l'admissibilité pour l'espace des paramètres  $[0, \infty)^N$  s'ensuit facilement. Un argument semblable sera utilisé pour la loi *a posteriori* de Pôlya

contraînte. Le traitement des contraintes sur des populations finies pose certaines problèmes techniques qu'il est difficile de régler. Par conséquent, nous supposons que la population est suffisamment grande comparativement à la taille de l'échantillon pour que la forme approximative de la loi *a posteriori* de Pôlya faisant intervenir la loi de Dirichlet est appropriée. Pour simplifier, nous supposons que la population  $U$  est infinie.

Nous supposons que pour tout  $j$  dans  $U$ ,  $(Y_j, X_j) = a_j$  pour un  $i$  dans  $\{1, \dots, k\}$ , où  $a_i = (a_{i1}^{(m+1)}, \dots, a_{iR^{m+1}}^{(m+1)})$  sont des vecteurs distincts dans  $R^{m+1}$  et où  $k$  peut être très grand. Autrement dit, les vecteurs  $(Y_j, X_j)$  ne peuvent prendre qu'un nombre fini de valeurs. Si  $p_i$  est la proportion de valeurs  $(Y_j, X_j)$  dans la population qui sont égales à  $a_i$ , pour  $i$  dans  $\{1, \dots, k\}$ , la moyenne de population de  $Y$  est

$$\sum_{i=1}^k p_i a_i^{(1)}.$$

Nous supposons que de l'information *a priori* est disponible au sujet des variables auxiliaires  $X^i := \{x_j^i \mid j \in U\}$  pour  $i$  dans  $\{1, \dots, m\}$ , ce qui donne lieu à des égalités et à des inégalités linéaires faisant intervenir les proportions  $p$  de la forme

$$A_j p = b_j \quad (4)$$

$$A_2 p \leq b_2 \quad (5)$$

où  $A_1, A_2$  sont les matrices de dimensions  $m_1 \times k$  et  $m_2 \times k$ , respectivement, et  $b_1, b_2$ , les vecteurs de dimensions appropriées. Dans ces conditions, par exemple, nous pouvons vouloir estimer

$$\pi(p) = \sum_{i=1}^l p_i a_i^{(1)},$$

sous les contraintes des équations 4 et 5 et où  $\sum_{i=1}^k p_i = 1$  avec  $p_i \geq 0$ , pour tout  $i$  dans  $\{1, \dots, k\}$ .

Considérons un échantillon  $s$  de taille  $n$  que, pour faciliter la notation, nous supposons être composé de  $n$   $a_i$  distincts. Soit  $a_s$  cet ensemble de valeurs. Nous posons alors que  $A_{1s}$  et  $A_{2s}$  sont les matrices de dimensions  $m_1 \times n$  et  $m_2 \times n$  qui sont simplement  $A_1$  et  $A_2$  restreintes

$$(7)$$

$$(6)$$

aux colonnes correspondant aux membres de  $a_s$ . Soit  $p_s$  la probabilité  $p$  restreinte aux membres de  $a_s$ . Alors, les contraintes sur la population données par les équations 4 et 5 se traduisent dans les contraintes suivantes

$$(7)$$

$$(6)$$

Une difficulté technique lorsque l'on prouve l'admissibilité sous des contraintes est que, même si la population satisfait les contraintes énoncées, il est toujours possible d'obtenir un échantillon qui ne les satisfait pas. Il existe plusieurs moyens de traiter ce genre de cas. L'un d'eux consiste à supposer que les contraintes sont incorrectes et de les ignorer. Cette tactique a été utilisée dans Nelson et Meeden (1998). Un autre consiste à utiliser l'information *a priori* pour augmenter l'échantillon de sorte qu'il satisfasse les contraintes. Cela peut être un exercice délicat et la réponse peut dépendre fortement de la façon dont l'échantillon est ajusté. Nous adoptons ici une troisième approche. Nous supposons que nous suivons un plan d'échantillonnage aléatoire simple et que notre information *a priori* doit être correcte. Dans ces conditions, il pourrait être sensé de rejeter tout échantillon qui ne satisfait pas les contraintes, puisqu'il s'agit clairement d'un échantillon non représentatif. Plus précisément, supposons que nous tirons un échantillon aléatoire simple de taille  $n$  de la population et que nous observons tous les  $x_j^i$  dans l'échantillon. Soit  $p_s^i = (p_1^i, \dots, p_k^i)$  les proportions des vecteurs possibles pour les  $x_j^i$  qui sont observés dans l'échantillon. L'élément  $p_s^i$  est nul chaque fois que le vecteur  $a_i = (a_{i1}^{(m+1)}, \dots, a_{iR^{m+1}}^{(m+1)})$  ne figure pas dans l'échantillon. Si  $p_s^i$  satisfait les équations 6 et 7, nous gardons l'échantillon, sinon nous l'écartons et essayons de nouveau. Nous donnons à ce plan le nom d'échantillonnage aléatoire sous contraintes. En pratique, pour des contraintes typiques, il ne sera jamais nécessaire d'écartier un échantillon. Bien que ce plan d'échantillonnage ne soit jamais utilisé, il ne s'agit pas d'une mauvaise approximation de ce qui est fait en réalité.

Plus formellement, soit  $Z_j$  le nombre de  $(Y_j, X_j)$  dans l'échantillon qui sont égales à  $a_i$ , pour  $i$  dans  $\{1, \dots, k\}$ , alors  $(Z_1, \dots, Z_k)$  est une loi Multinomiale( $n, p_1, \dots, p_k$ ) où les valeurs des paramètres appartiennent à

$$P := \left\{ (p_1, \dots, p_k) \mid A_1 p = b_1, A_2 p \leq b_2, \sum_{i=1}^k p_i = 1, \text{ et } p_i \geq 0 \forall i \in \{1, \dots, k\} \right\}. \quad (8)$$

variantes, telles que la loi *a posteriori* de Polya contraire, ne dépendent pas directement de l'échantillonnage aléatoire simple, de l'échantillonnage aléatoire stratifié ou de tout autre plan d'échantillonnage. Leur caractère approprié et leur performance dépendent de la concordance entre la structure sous-jacente de la population et celle spécifiée dans la loi prédictive choisie.

La notion fondamentale qui étaye la PPC est que l'on devrait utiliser l'échantillon et l'information auxiliaire disponibles pour simuler des répliques complètes, représentatives, de la population. Dans de simples exemples tels que ceux donnés plus haut, nous voyons que son estimateur ponctuel devrait avoir d'excellentes propriétés fréquentistes pour une vaste classe de plans d'échantillonnage et que la performance de son estimateur par intervalle est adéquate si la taille d'échantillon n'est pas trop faible. Cela signifie-t-il que la PPC peut ajuster automatiquement de « mauvais » échantillons afin d'obtenir de bonnes estimations? Pas vraiment, puisque si nous avons un très mauvais échantillon, c'est-à-dire un échantillon qui est médiocrement compatible avec l'information *a priori* connue, deux choses indésirables peuvent se produire. Premièrement, des échantillons très déséquilibrés ou biaisés induiront un certain biais dans l'estimation ponctuelle. Deuxièmement, ils restreindront sévèrement les valeurs possibles de  $p$  sous la PPC et donneront lieu à une variance *a posteriori* trop faible, ce qui donnera des intervalles estimés qui sont trop courts et qui couvriront insuffisamment le paramètre d'intérêt. Dans les situations plus compliquées, il faudra poursuivre l'étude afin de découvrir quand l'utilisation de la loi *a posteriori* de Polya contraire peut être avantageuse.

## 7. Remarques finales

L'un des inconvénients des méthodes fréquentistes classiques est que chaque problème demande sa propre solution. Estimer la médiane de population de  $y$  quand la moyenne de population  $x$  est connue n'est pas le même problème qu'estimer la moyenne de  $y$  quand la moyenne de  $x$  est connue. En outre, si la moyenne de population de  $x$  n'est pas connue exactement et que l'on sait seulement qu'elle est comprise dans un certain intervalle de valeurs, les méthodes fréquentistes classiques ne peuvent pas utiliser cette information. L'une des forces d'une approche *a posteriori* qui combine raisonnablement l'échantillon avec l'information *a priori*, les inférences peuvent être faites pour de nombreux paramètres de population d'intérêt en simulant simplement des répliques complètes de la population.

Ici, nous avons soutenu que la loi *a posteriori* de Polya contraire est une méthode raisonnable d'introduction d'information *a priori* objective au sujet des variables

auxiliaires dans une approche bayésienne non informative de l'échantillonnage d'une population finie. Les estimateurs ponctuels résultants ont une justification bayésienne séquentielle qui garantit leur admissibilité. Leurs intervalles de crédibilité à 95 % sont habituellement des intervalles de confiance à 95 % approximatifs et ils fournissent des réponses raisonnables pour des problèmes pour lesquels aucune méthode fréquentiste classique n'est disponible. Cela démontre une force importante de la loi *a posteriori* de Polya. Une fois que l'on peut simuler des répliques raisonnables de la population complète, l'inférence pour toute une gamme de problèmes devient simple. Par contre, on doit utiliser des méthodes MCMC pour leur calcul. Tous nos calculs ont été effectués en R (R Development Core Team 2005). Deux des auteurs ont diffusé récemment un projeté R appelé Polypost qui permet à d'autres d'utiliser facilement nos méthodes. Ici, nous nous sommes limités à des échantillons de taille inférieure à 100, uniquement pour des raisons pratiques, afin de pouvoir exécuter nos simulations en un temps raisonnable. En pratique, pour un échantillon spécifique plus grand, il suffit d'exécuter une chaîne plus longue. Puis, on peut utiliser certains diagnostics standard pour décider si elle semble ou non avoir convergé.

## Annexe

### Une preuve d'admissibilité

La justification théorique fondamentale pour les estimateurs ponctuels découlant de la loi *a posteriori* de Polya est qu'ils sont admissibles. Les preuves d'admissibilité s'appuient sur la nature bayésienne séquentielle de la loi *a posteriori* de Polya. Nous présentons ici une preuve pour les estimateurs ponctuels fondés sur la loi *a posteriori* de Polya contraire.

Dans ces arguments bayésiens séquentiels, nous sélectionnons une séquence finie de sous-ensembles disjoints de l'espace des paramètres, où l'ordre est important. Nous définissons une loi *a priori* différente pour chacun de ces sous-ensembles. Ensuite, nous trouvons l'estimateur de Bayes pour chaque point échantillon qui reçoit une probabilité positive sous la première loi *a priori*. Puis, nous trouvons l'estimateur de Bayes pour tout point échantillon qui reçoit une probabilité positive sous la deuxième loi *a priori* et qui n'est pas considéré sous la première loi *a priori*. Ensuite, nous considérons la troisième loi *a priori*, et ainsi de suite. Pour un point échantillon particulier, la valeur de l'estimation bayésienne séquentielle est la valeur de l'estimateur bayésien provenant de l'étape à laquelle ce point a été considéré. La nature bayésienne séquentielle de la loi *a posteriori* de Polya explique ses propriétés que l'on peut paradoxalement. Pour un échantillon donné, elle se



L'information supplémentaire, ce qui aboutit à un estimateur ponctuel amélioré, mais la couverture des estimations des intervalles reste insuffisante.

Tableau 8

Résultats des simulations pour les deux populations de la figure 1 quand on estime le total de 400 échantillons qui ont été sélectionnés en utilisant PPS(x) comme plan d'échantillonnage. Notons que absent est la valeur absolue de l'écart entre les valeurs estimée et réelle du total de population

Méthode	Estimation	Erreur Longueur	Fréquence	couverture
Pour la population 1 avec le total = 118 210,2				
HT	118 803,1	8 095,3	38 696,6	0,905
REG	116 838,1	3 355,3	14 136,4	0,905
PPC	117 515,7	3 277,3	14 330,7	0,905
pour une taille d'échantillon de 60				
HT	119 139,2	5 395,6	28 233,3	0,952
REG	117 041,4	2 213,2	9 561,0	0,910
PPC	118 041,4	2 195,3	11 836,5	0,938
Pour la population 2 avec le total = 11 648,7				
HT	11 737,2	783,5	4 012,0	0,945
REG	11 800,3	179,7	533,0	0,745
PPC	11 689,9	122,4	535,4	0,900
REG2	11 660,0	97,2	382,3	0,862
PPC2	11 689,9	122,4	535,4	0,900
pour une taille d'échantillon de 60				
HT	11 774,2	564,8	2 908,2	0,955
REG	11 795,8	155,2	373,1	0,635
PPC	11 647,9	80,4	524,4	0,978
REG2	11 663,1	66,7	266,2	0,895
PPC2	11 651,2	88,4	523,6	0,962
PPCbd	11 644,6	83,9	552,1	0,978

Pour la deuxième population, nous avons exécuté un troisième ensemble de simulations en utilisant PPS(x) comme plan pour les tailles d'échantillon de 30 et de 60. Dans ce cas, nous avons supposé que la moyenne de population de  $x$  était contenue dans l'intervalle (4,45, 5,53). Ces valeurs correspondent au 45° et au 65° quantiles de la population  $x$ . La moyenne de cette population est égale à 5,02. Les résultats sont présentés au tableau 8 sous l'étiquette PPCbd. Nous voyons que les résultats sont fort semblables à ceux obtenus en supposant que la moyenne de population de  $x$  était connue.

Les trois estimateurs utilisent l'information contenue dans la variable auxiliaire  $x$ , mais l'estimateur HT est le seul qui dépend de la connaissance du plan d'échantillon-nage. Comme nous l'avons mentionné, il est bien connu que les estimateurs bayésiens ne font pas entrer les probabilités d'échantillonnage dans les calculs. Dans les exemples présentés ici, nous voyons que l'estimateur PPC utilise efficacement l'information contenue dans la variable auxiliaire. En général, la loi *a posteriori* de Polya et ses

partie la même, quoique la différence entre REG et PPC soit nettement plus petite pour la deuxième population. Par exemple, quand la taille d'échantillon est égale à 30, l'erreur absolue moyenne et la fréquence de couverture sont égales à 131,9 et 0,875, respectivement, pour REG comparativement à 124,3 et 0,908 pour PPC. Quand la taille d'échantillon est égale à 60, les chiffres pour REG sont 88,4 et 0,905 comparativement à 90,1 et 0,958 pour l'estimateur PPC. Les longueurs moyennes de leurs intervalles sont égales à 384 et à 560, respectivement.

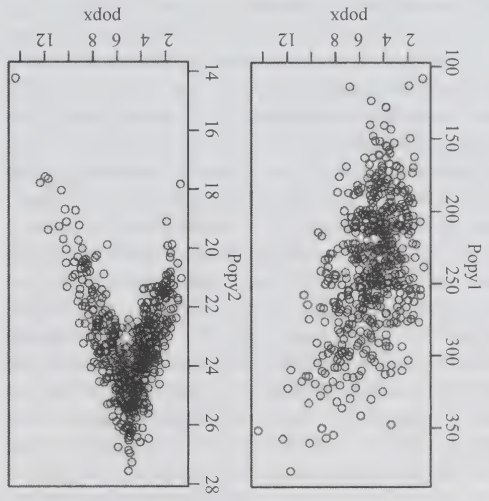


Figure 1 Traces des deux populations utilisées dans les simulations au tableau 8. Les corrélations pour les deux populations sont égales à 0,47 et -0,22 et leurs totaux sont 118 210 et 11 648,7

Pour la deuxième population, nous avons réalisé un deuxième ensemble de simulations en utilisant le plan PPS(x) pour les tailles d'échantillon de 30 et de 60. Cette fois-ci, nous avons supposé que les moyennes de population de  $x$  et  $x^2$  étaient connues. Nous avons alors comparé l'estimateur PPC qui intègre les contraintes sur  $x$  ainsi que sur  $x^2$  à l'estimateur par la régression reposant sur l'hypothèse d'une fonction quadratique de  $x$  comme modèle. Ces estimateurs sont désignés par PPC2 et REG2 dans le tableau. À première vue, il pourrait paraître étonnant que les résultats pour PPC et PPC2 soient essentiellement les mêmes. Cependant, si l'on y réfléchit, ce résultat est celui auquel on devrait s'attendre. La loi *a posteriori* de Polya contraint la simule des répliques entières de la population, qui sont « équilibrées » en ce qui concerne  $x$ , autrement dit qui concordent avec sa moyenne connue de population. La contrainte supplémentaire qu'une réplique simulée de la population doive aussi être « équilibrée » par rapport à  $x^2$  ajoute peu d'information. Par ailleurs, avec un modèle raisonnable, l'estimateur par la régression peut exploiter

moyenne de 0,91. Donc, si l'on s'en tient à l'erreur absolue, la performance de l'estimateur de la loi *a posteriori* de Polya contrainte était supérieure d'environ 10 %, en moyenne, à celle de l'estimateur de Chen et Qin.

Supposons maintenant que la médiane de population de  $x$  est connue. Pour simplifier, posons qu'aucune des valeurs réelles  $n$  est égale à la médiane de population de  $x$ . Soit  $n_i$  le nombre d'unités de l'échantillon dont les valeurs de  $x$  sont inférieures à la médiane de population connue de  $x$ . Alors,  $n_u = n(s) - n_i$  est le nombre d'unités de l'échantillon qui sont situées de l'autre côté de la médiane connue. Soit  $p_1 = (d_1, \dots, d_{n_i})$  et  $p_u = (d_{n_i+1}, \dots, d_n)$  les deux vecteurs de probabilités. Intuitivement, une loi *a posteriori* raisonnable, étant donné l'échantillon et la médiane de population connue, serait que  $p_i$  et  $p_u$  suivent des lois de Dirichlet indépendantes dont tous les paramètres sont égaux à un et affectés chacun d'un poids égal à un demi, afin que leur somme soit égale à un. Il découle du théorème prouvé en annexe que, sous notre plan d'échantillonnage, ces lois *a posteriori* sont des bayésiennes séquentielles. Notons que sous ces lois *a posteriori*, les valeurs espérées des proportions affectées à chaque unité de l'échantillon sont les poids affectés à l'échantillon par Chen et Qin. Cela prouve l'admissibilité de leur estimateur de la médiane de population et, conséquemment, de l'estimateur de Kuk et Mak. Les résultats des simulations montrent que les intervalles de crédibilité à 95 % de cette loi *a posteriori* de Polya contraignent couvrent environ 95 % du temps, sauf dans un cas particulier. Si la taille d'échantillon est faible et que  $y$  et  $x$  sont fortement corrélés, les médianes des populations simulées sous la loi *a posteriori* de Polya contraignent ne varient pas suffisamment et les intervalles résultants sont trop courts et leur fréquence de couverture pourrait être considérablement inférieure à 95 %.

Cette relation étroite entre l'approche de vraisemblance étonnante si l'on souligne que, dans le cas non contraint, la séquence des lois *a priori* menant à la loi *a posteriori* de Polya peut être utilisée pour prouver l'admissibilité de l'estimateur du maximum de vraisemblance pour le vecteur de probabilités d'une loi multinomiale.

## 6. Autres plans d'échantillonnage

Tous les résultats des simulations présentés jusqu'à présent ont été obtenus en utilisant un plan d'échantillonnage aléatoire simple (stratifié) sans remise (EAS). L'un des examinateurs d'une version antérieure du manuscrit soumettait savoir quelle mesure le comportement des estimateurs fondés sur la loi *a posteriori* de Polya contraignent dépendent de l'utilisation de ce plan d'échantillonnage. La réponse est qu'il existe une certaine dépendance, mais dépendent de l'utilisation de ce plan d'échantillonnage. La

Nous avons constaté, sous l'EAS, que l'estimateur fondés sur la loi *a posteriori* de Polya contrainte (PPC) se comporte de façon fort semblable à l'estimateur par la régression (REG). Formellement, l'estimateur par la régression dépend uniquement de la connaissance de la moyenne de population de la variable auxiliaire. Ces propriétés sont habituellement étudiées sous échantillonnage aléatoire simple et l'estimateur de sa variance n'est valide que pour les grands échantillons.

Pour un plan général, l'estimateur d'Horvitz-Thompson (HT) est fréquemment utilisé. Il est sans biais, mais le calcul des probabilités d'inclusion exactes peut être difficile. Il en est ainsi, par exemple, si l'échantillonnage est fait avec des probabilités de sélection proportionnelles à la taille pour une variable auxiliaire  $x_i$ , disons  $PPS(x_i)$ . En pratique, on suppose simplement que la probabilité d'inclusion d'une unité est proportionnelle à sa valeur de  $x_i$  et que l'estimateur résultant sera approximativement sans biais.

Nous avons exécuté plusieurs études par simulation en vue de comparer ces trois méthodes pour l'estimation d'un total de population. Afin d'abréger, nous présentons les résultats de deux de ces études. Dans le cadre de celles-ci, nous avons construit deux populations de taille égale à 500. La variable auxiliaire est la même dans les deux populations et est un échantillon aléatoire tiré d'une loi gamma dont le paramètre de forme est égal à 5 et le paramètre d'échelle, à 1. Les tracés des deux populations sont présentés à la figure 1. Nous ne nous-entendons pas qu'en pratique, on utiliserait vraisemblablement l'estimateur par la régression dans la deuxième population. Il est présenté ici simplement pour illustrer ce qui peut arriver.

Pour chaque population, nous avons tiré 400 échantillons aléatoires de tailles égales à 30 et à 60 sous deux plans d'échantillonnage différents, à savoir  $PPS(x_i)$  et  $PPS(1/(x_i + 5))$ . Nous avons supposé que la moyenne de population de  $x$  était connue. Pour chaque échantillon, nous avons calculé les trois estimations du total de population. Les résultats pour le premier plan d'échantillonnage sont donnés au tableau 8. Nous voyons que l'estimateur PPC l'emporte clairement sur les autres. La fréquence de couverture observée de l'estimateur HT de l'intervalle est celle qui s'approche le plus du niveau nominal de 95 %. Mais l'intervalle est ridiculement long, parce que les inverses des probabilités d'inclusion varient fortement. Pour la première population, qui est approximativement linéaire, les estimateurs REG et PPC se comportent de la même manière. Cependant, pour la deuxième population, PPC donne de meilleurs résultats que REG. Son seul défaut ici est que sa couverture est insuffisante pour la plus petite taille d'échantillon. Sous  $PPS(1/(x_i + 5))$ , l'histoire est en grande



exemple, supposons que la moyenne d'échantillon est inférieure à la moyenne de population connue des valeurs de  $x$ . Alors, nous choisissons pour les unités échantillonnées des poids positifs, dont la somme est égale à 1, de sorte que la moyenne des valeurs  $x_j$  sous la loi de probabilité donnée par les poids satisfasse la contrainte connue. Bien qu'il ne soit pas possible de trouver explicitement ces poids, il est facile de les calculer. Quand nous estimons la moyenne de population de  $y$ , nous obtenons un estimateur qui a été mentionné pour la première fois dans Hartley et Rao (1968) et dont l'équivalence asymptotique à l'estimateur par la régression a été démontrée. Si la médiane de population de  $x$  est connue, nous attribuons aux unités de l'échantillon dont la valeur est inférieure à cette médiane des poids égaux dont la somme est égale à 0,5 et nous en faisons de même pour les unités échantillonnées dont les valeurs de  $x$  sont supérieures. Quand nous estimons la médiane de population, l'estimateur résultant est l'un de ceux proposés par Kuk et Mak (1989).

Un avantage de la loi *a posteriori* de Polya contrainte et plus généralement de l'approche bayésienne tient au fait qu'il est facile d'estimer un grand nombre de paramètres de population, outre la moyenne, sans devoir élaborer une nouvelle théorie ou de nouvelles méthodes. Étant donné une réplique simulée de la population complète qui satisfait les contraintes, on calcule simplement les paramètres de population d'intérêt. Ensuite, on utilise ces valeurs simulées exactement de la même façon que quand on estime la moyenne.

Pour comparer l'estimateur de Chen et Qin de la médiane de population de  $y$  à l'estimateur de la loi *a posteriori* de Polya contrainte lorsque la moyenne de population de  $x$  est connue, nous avons construit huit populations différentes. Pour la moitié de ces populations, on pourrait s'attendre à ce que l'estimateur par la régression donne de bons résultats pour l'estimation de la moyenne de population, tandis que l'autre moitié n'a pas satisfait les hypothèses habituelles de modélisation d'une superpopulation associées à l'estimateur par la régression. Pour chaque population, nous avons tiré 500 échantillons aléatoires de taille égale à 30 et à 50, sous la contrainte que l'échantillon contienne les valeurs de  $x$  supérieures et inférieures à la moyenne connue. Dans tous les cas, les deux estimateurs utilisant l'information *a priori* ont donné de meilleurs résultats que la médiane d'échantillon. Ces résultats concordent avec les résultats de simulation de Chen et Qin. Nous avons calculé l'erreur absolue moyenne pour les deux estimateurs utilisant la contrainte de moyenne. Dans chacun des 16 ensembles distincts de simulation, nous avons alors calculé le ratio de l'erreur absolue de l'estimateur de la loi *a posteriori* de Polya contrainte à celle de l'estimateur de Chen et Qin. L'intervalle de ces 16 valeurs allait de 0,85 à 1,00 avec une

Tableau 6  
Tailles des strates et tailles des échantillons pour les données de la Veterans Administration

Bonne	Passable	Mauvaise
F	33(20)	155(10)
H	890(30)	493(20)
	492(10)	

Nous avons comparé trois estimateurs différents du score PCS moyen pour cette population de 2 500 personnes, à savoir la moyenne d'échantillon calculée sans tenir compte de la stratification, l'estimateur stratifié habituel en supposant que les tailles de strate sont connues et un estimateur fondé sur la loi *a posteriori* de Polya contrainte en supposant que les totaux de marge des lignes et des colonnes du tableau 6 étaient connus, de même que l'âge moyen des membres de la population. La corrélation de population entre le score PCS et l'âge est de -0,22. Les corrélations du score PCS avec le sexe et avec la catégorie d'état de santé fondée sur la comorbidité sont égales à -0,13 et -0,28. Les résultats du tableau 7 nous montrent que l'estimateur fondé sur la loi *a posteriori* de Polya contrainte donne à peu près le même résultat que l'estimateur stratifié et que tous deux sont un peu meilleurs que la moyenne d'échantillon. Afin de calculer l'estimateur fondé sur la loi *a posteriori* de Polya contrainte, nous avons produit des chaînes de Markov de longueur égale à 7 000 000.

Tableau 7  
Résultats de l'estimation du score PCS d'après les données de la Veterans Administration. L'estimateur de la loi *a posteriori* de Polya contrainte suppose que les totaux de ligne et de colonne sont connus, ainsi que l'âge moyen des membres de la population

Méthode	Estimation ponctuelle	Estimation ponctuelle	Erreur absolue	Borne inférieure	Longueur moyenne	Fréquence de couverture
Moyenne	37,235	1,040	34,907	4,650	0,938	0,948
Stratification	36,648	0,925	34,322	4,651	0,948	0,958
Polya	36,644	0,925	34,344	4,605	0,958	

### 5. Relation avec les méthodes de vraisemblance empiriques

Nous allons maintenant examiner certaines méthodes fréquentistes pour la résolution de problèmes comportant des contraintes et discuter de leur relation avec la loi *a posteriori* de Polya contrainte.

Chen et Qin (1993) ont étudié une approche de vraisemblance empirique de l'estimation par sondage lorsque l'on dispose d'information *a priori* au sujet d'une caractéristique auxiliaire. Pour construire des estimateurs après que l'échantillon a été observé, les unités échantillonnées sont pondérées de façon à refléter l'information *a priori*. Par



**Tableau 5**  
Résultats des simulations pour l'exemple catégorique quand les totaux de marge sont supposés connus et que l'on sait que la moyenne de la variable auxiliaire est comprise entre ses 45<sup>e</sup> et 65<sup>e</sup> quantiles connus

Méthode	Estimation	Estimation ponctuelle	Borne inférieure	Borne supérieure	Fréquence moyenne	Couverture de
Moyenne	43,463	0,840	40,789	5,348	0,978	
Stratification	43,948	0,865	41,825	4,245	0,948	
Polya	43,519	0,829	41,555	4,029	0,938	

4.3 Un exemple

À la présente section, nous considérons les données provenant de la Veterans Health Administration. En 1998, le VA Upper Midwest Health Care Network a administré une enquête sur l'état fonctionnel des anciens combattants utilisateurs des établissements de la VA dans le réseau (Singh, Borowsky, Nugent, Murdoch, Zhao, Nelson, Petzel et Nichol 2005). Les anciens combattants admissibles à cette enquête étaient ceux qui avaient reçu des soins au service des consultations externes ou qui avaient été hospitalisés entre octobre 1997 et mars 1998 dans l'un des cinq établissements de la VA du réseau. En plus des mesures démographiques de base, telles que l'âge et le sexe, la principale composante de l'enquête était le questionnaire SF36-V (Kazis, Miller, Clark, Skinner, Lee, Rogers, Spito, Payne, Finck, Selim et Linzer 1998). Ce questionnaire sur la qualité de la vie liée à l'état de santé comprend huit sous-échelles pour le fonctionnement physique, les limitations des rôles dus à des problèmes physiques, la douleur corporelle, l'état de santé général, l'énergie/vitalité, le fonctionnement social, les limitations des rôles dues à des problèmes émotionnels et la santé mentale. Ces échelles sont combinées pour produire les scores sommatifs de la composante physique (PCS) et de la composante mentale (MCS). L'état de santé est d'autant meilleur que ces scores sont élevés. Des données administratives de la VHA mesurant les principaux états comorbides présents l'année qui a précédé l'enquête ont également été recueillies.

À partir de la population de l'un des cinq établissements, nous avons sélectionné une sous-population comprenant toutes les femmes et un sous-ensemble aléatoire des hommes pour créer une population de 2 500 personnes. Pour les besoins du présent exemple, les nombres de comorbidités ont été regroupés en trois catégories correspondant à une bonne santé, une santé passable et une mauvaise santé. Puis, nous avons tiré à partir de la population 200 échantillons aléatoires stratifiés de taille égale à 100. Les tailles de strate, ainsi que les tailles d'échantillon sont présentées au tableau 6. Notre plan d'échantillonnage prévoyait le sur-échantillonnage des femmes. Ce genre de plan d'échantillonnage non équilibré est utilisé fréquemment en pratique.

Provenant de chaque cellule. Pour chaque échantillon, nous avons calculé la moyenne d'échantillon et l'estimation par stratification, en supposant que les effectifs de population réels des cellules étaient connus, ainsi que les intervalles de confiance à 95 % correspondants. Nous avons également calculé les estimations à partir de la loi *a posteriori* de Polya contrairement, ainsi que leurs intervalles de crédibilité à 0,95. Nous avons obtenu ces dernières estimations à partir des derniers 4 000 000 de points d'une chaîne de Markov de longueur égale à 4 100 000. Les résultats des simulations sont présentés dans les tableaux 2 à 5. Ils révèlent que les estimations d'après la loi *a posteriori* de Polya contrairement fondée sur les totaux de marge et une moyenne ou une médiane connue ou un intervalle connu autour de la moyenne sont meilleures que les estimations stratifiées ne sont meilleures que les estimations stratifiées ne sont meilleures que les estimations d'après la loi *a posteriori* de Polya contrairement fondée sur les totaux de marge connus.

**Tableau 2**  
Résultats des simulations pour l'exemple catégorique quand seuls les totaux de marge sont supposés connus

Méthode	Estimation	Estimation ponctuelle	Borne inférieure	Borne supérieure	Fréquence moyenne	Couverture de
Moyenne	43,805	0,919	41,107	5,396	0,976	
Stratification	44,355	0,846	42,259	4,191	0,940	
Polya	43,909	0,896	41,863	4,197	0,922	

**Tableau 3**  
Résultats des simulations pour l'exemple catégorique quand les totaux de marge et la moyenne de la variable auxiliaire sont supposés connus

Méthode	Estimation	Estimation ponctuelle	Borne inférieure	Borne supérieure	Fréquence moyenne	Couverture de
Moyenne	43,804	0,922	41,063	5,482	0,964	
Stratification	44,399	0,862	42,272	4,256	0,942	
Polya	44,506	0,810	43,257	2,497	0,960	

**Tableau 4**  
Résultats des simulations pour l'exemple catégorique quand les totaux de marge et la médiane de la variable auxiliaire sont supposés connus

Méthode	Estimation	Estimation ponctuelle	Borne inférieure	Borne supérieure	Fréquence moyenne	Couverture de
Moyenne	43,439	0,877	40,783	5,312	0,986	
Stratification	43,927	0,884	41,804	4,244	0,940	
Polya	43,784	0,785	42,032	3,640	0,920	

Tableau 1  
Résultats des simulations pour l'exemple stratifié où la médiane de la première variable auxiliaire et la moyenne de la seconde sont connues

Méthode	Estimation	Intervalle de confiance ou de pontuelle	Estimation ponctuelle	Erreur absolue	Erreur relative	Longueur	Fréquence de couverture
Moyenne	47,978	4,821	36,44	23,09	1,000		
Stratification	43,995	2,072	38,22	10,35	0,942		
Polya	43,355	1,516	40,19	6,75	0,936		

4.2 Variables auxiliaires catégoriques

Supposons que les éléments d'une population de taille connue  $N$  sont associés aux éléments de  $k$  variables auxiliaires catégoriques. Pour simplifier, nous posons que  $k = 2$ , mais la théorie s'applique à plus de deux variables catégoriques. Si une variable auxiliaire prend  $r$  valeurs distinctes et que l'autre prend  $c$  valeurs distinctes, elles permettent de classer les éléments de la population dans un tableau à double entrée contenant  $r \times c$  cellules. Soit  $N_{ij}$  le nombre d'éléments de la population qui appartiennent à la cellule  $ij$ , pour  $i$  dans  $\{1, \dots, r\}$  et  $j$  dans  $\{1, \dots, c\}$ , alors  $\sum_{i=1}^r \sum_{j=1}^c N_{ij} = N$ . Si les  $N_{ij}$  sont connus et que  $s$  est un échantillon aléatoire contenant  $n_{ij}$  éléments provenant de la cellule  $ij$ , alors une bonne estimation de la moyenne de

où  $\bar{y}_s^{ij}$  est la moyenne des  $n_{ij}$  éléments provenant de la cellule  $ij$  dans l'échantillon. Il s'agit de l'estimateur stratifié habituel, où les cellules du tableau sont considérées comme étant les strates.

Un problème plus ardu est celui de l'estimation de la moyenne de population quand les dénombrements,  $N_{ij}$ , ne sont pas connus, mais que les dénombrements totaux de marge le sont. Soit  $N_{.j} = \sum_{i=1}^r N_{ij}$  les totaux de marges, pour  $i$  dans  $\{1, \dots, r\}$  et  $N_{i.} = \sum_{j=1}^c N_{ij}$  les totaux de marges des colonnes, pour  $j$  dans  $\{1, \dots, c\}$ . Dans de tels cas, un moyen d'estimer la moyenne de population consiste à appliquer la méthode fréquentiste appelée calage ou méthode itérative du quotient. Selon cette méthode, étant donné un échantillon  $s$ , l'estimateur est donné par  $\sum_{i \in s} w_k y_k$ , où les  $w_k$  ne sont pas les poids de sondage, mais de nouveaux poids attribués aux unités de l'échantillon. Un bon ensemble de poids doit satisfaire deux conditions. La première est que les poids doivent préserver les totaux de marge connus, par exemple,  $\sum_{k \in s(j)} w_k = N_{.j}$  (où  $s(j)$  est la portion de l'échantillon se trouvant dans la  $j^{\text{e}}$  colonne du tableau à double entrée. La deuxième est que les poids devraient, dans un certain sens, être proches des poids d'échantillonnage  $1/\pi_k$ , où  $\pi_k = P(k \in s)$ . Suivant

la fonction utilisée pour mesurer la distance, divers estimateurs par calage peuvent être obtenus. Bien qu'il s'agisse d'une idée raisonnable, la sélection de la bonne mesure de distance et l'obtention d'une estimation raisonnable de la variance pour l'estimateur résultant n'ont pas de réponse fréquentiste standard. Pour des détails, voir Deville et Särndal (1992). La loi *a posteriori* de Polya donne une approche alternative de résolution du problème, puisque l'information fournie par les totaux de marge connus détermine un ensemble de contraintes linéaires sur les poids aléatoires attribués aux unités échantillonnées. S'il existe des variables auxiliaires continues pour lesquelles nous disposons d'information *a priori*, nous pouvons ajouter des contraintes supplémentaires. Pour voir comment cette approche fonctionnerait en pratique, nous avons envisagé l'exemple simple de deux variables dichotomiques, afin que chaque unité puisse être classée dans une cellule d'un tableau  $2 \times 2$ , et d'une troisième caractéristique auxiliaire continue. Nous avons émis l'hypothèse de quatre niveaux distincts d'information *a priori*.

1. Les totaux de marge pour le tableau à double entrée sont connus.
  2. Les totaux de marge et la moyenne de la variable auxiliaire continue sont connus.
  3. Les totaux de marge et la médiane de la variable auxiliaire sont connus.
  4. Les totaux de marge sont connus et l'on sait que la moyenne de la variable auxiliaire continue est comprise entre deux bornes. Nous avons choisi les 45<sup>e</sup> et 65<sup>e</sup> quantiles de sa population de valeurs pour spécifier ces bornes.
- Pour chaque cas, nous avons formé une population en utilisant le modèle qui suit, où toutes les variables aléatoires sont indépendantes.

Cellule 1,1  $x_i \sim \text{gamma}(8, 1)$ ,  $e_i \sim \text{normale}(0, .7^2)$  et  $y_i = 25 + 3x_i + e_i$  pour  $i$  dans  $\{1, \dots, 150\}$ .  
Cellule 1,2  $x_i \sim \text{gamma}(10, 1)$ ,  $e_i \sim \text{normale}(0, .7^2)$  et  $y_i = 25 + 3x_i + e_i$  pour  $i$  dans  $\{1, \dots, 350\}$ .  
Cellule 2,1  $x_i \sim \text{gamma}(6, 1)$ ,  $e_i \sim \text{normale}(0, .4^2)$  et  $y_i = 25 + 2x_i + e_i$  pour  $i$  dans  $\{1, \dots, 250\}$ .  
Cellule 2,2  $x_i \sim \text{gamma}(4, 1)$ ,  $e_i \sim \text{normale}(0, .4^2)$  et  $y_i = 25 + 2x_i + e_i$  pour  $i$  dans  $\{1, \dots, 250\}$ .  
Pour chacun des cas, nous avons généré une population et tiré 500 échantillons aléatoires de taille 80 avec 20 unités



estime les quantités plutôt que les moyennes. Voir, par exemple, Efron et Tibshirani (1993). Une autre approche qui peut donner de bons résultats consiste à exécuter la chaîne pendant une longue période et à n'utiliser que chaque  $m^e$  point, où  $m$  est un grand nombre entier. La méthode est inefficace, mais elle peut donner de bonnes réponses lorsqu'on veut trouver un intervalle de crédibilité à 95 % pour la médiane.

#### 4. Applications

À la présente section, nous montrons comment divers types d'information partielle au sujet des variables auxiliaires peuvent être intégrés dans l'estimation des paramètres quand on utilise la loi *a posteriori* de Polya contraire. Dans de nombreux cas, l'information *a priori* utilisée dans l'estimation selon la loi *a posteriori* de Polya contraire ne peut pas l'être dans les méthodes fréquentistes classiques.

##### 4.1 Stratification

La stratification est une forme d'information *a priori* qui est utilisée couramment en échantillonnage d'une population finie. Nous notons que l'estimateur stratifié habituel peut être considéré comme découplant de lois *a posteriori* de Polya indépendantes dans chaque strate. Des renseignements détaillés figurent dans Vardeman et Meeden (1984). Quand, en plus de la stratification, il existe une variable auxiliaire, on peut obtenir une bonne estimation de la moyenne de population en combinant les estimations produites à l'aide de l'estimateur par régression dans chaque strate. Pour plus de précisions, voir Cochran (1976). Si l'on ne connaît que la moyenne de population de la variable auxiliaire, sous les approches classiques, il est difficile de combiner cette information avec la stratification, à moins que l'on émette l'hypothèse d'un modèle commun à toutes les strates. La Polya contraire peut intégrer les deux types d'information susceptibles d'améliorer les estimations, pourtant elle ne nécessite pas l'hypothèse d'un modèle commun.

Pour le démontrer, nous avons créé une population stratifiée de 900 valeurs comportant trois strates. Les tailles des strates étaient 300, 200 et 400. Nous avons considéré deux variables auxiliaires, disons  $x_1$  et  $x_2$ . Dans la première strate, les  $x_{1,i}$  étaient un échantillon aléatoire tiré d'une loi gamma (10,1) et les  $x_{2,i}$  étaient un échantillon aléatoire tiré d'une loi gamma (2,1). Dans la deuxième strate, les  $x_{1,i}$  et les  $x_{2,i}$  étaient tirés des lois gamma (5,1) et gamma (3,1), respectivement. Dans la troisième strate, les  $x_{1,i}$  et les  $x_{2,i}$  étaient tirés des lois gamma (15,1) et gamma (7,1), respectivement.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Dans cet exemple, les estimations au moyen de la loi de Polya contraire ont été obtenues en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.

Polya contraire ont été obtenus en utilisant des chaînes de Markov ayant une longueur de 4 000 000 après avoir écarté les 100 000 premiers points.



population de  $x^j$ , alors la contrainte pour la population

stimulée devient

$$\sum_{i=1}^j p^f_j(x^{t_i}_2 - x^3_{t_j}) \leq 0.$$

Donc, étant donné une famille de contraintes de popula-

tion fondées sur l'information *a priori* et un échantillon,

nous pourrions représenter les contraintes correspondantes sur le paramètre  $p$  simulé par deux systèmes d'équations

$$(2) \qquad A_{1,s} p = b_1$$

$$(3) \qquad A_{2,s} p \leq b_2$$

où  $A_{1,s}$  et  $A_{2,s}$  sont des matrices de dimensions  $m_1 \times n(s)$  et  $m_2 \times n(s)$ , et  $b_1$  et  $b_2$  sont des vecteurs ayant les dimen-

sions appropriées. Cela généralise l'argument menant à l'équation 1.

Nous supposons l'échantillon tel que le sous-ensemble

du simplexe qu'il définit par les équations 2 et 3 n'est pas vide. Pour un tel échantillon, l'approximation asymptotique de la loi *a posteriori* de Pólya contrainte impose une loi uniforme sur ce sous-ensemble du simplexe. Avant d'aborder la question de la simulation à partir de cette loi, nous notons qu'elle possède une justification théorique. Il peut lui être donné une justification bayésienne séquentielle qui garantit qu'elle produira des méthodes admissibles. Des renseignements détaillés figurent en annexe.

### 3.3 Calculs

Soit  $P$  le sous-ensemble du simplexe qui est défini par les équations 2 et 3.  $P$  est un polytope de dimension non pleine. Nous aimerions produire des observations indépendantes à partir de la loi uniforme sur  $P$ . Malheureusement, nous ne savons pas comment le faire. Nous utilisons plutôt des méthodes de Monte Carlo par chaînes de Markov (MCMC) pour produire des échantillons dépendants.

En particulier, nous utilisons l'algorithme de Metropolis-Hastings qui dépend de l'emploi d'une chaîne de Markov pour produire une séquence dépendante de valeurs aléatoires pour  $p \in P$ . Le processus se déroule comme suit. Nous commençons par trouver un point de départ en  $p_0$  dans l'intérieur relatif de  $P$ . Il s'agit de l'étape 1 énoncée plus bas. Ensuite, nous choisissons une direction aléatoire  $d$  dans  $P$ , ce qui pose un peu de difficulté parce que la dimension de  $P$  est strictement inférieure à  $n(s) - 1$ . Cela est accompli aux étapes 2 et 3 décrites plus bas. Ensuite, nous trouvons le segment de droite qui est l'intersection de la droite passant par  $p_0$  dans la direction  $d$  avec  $P$ . Il s'agit de l'étape 4 énoncée plus bas. Puis, nous choisissons un point au hasard à partir de la loi uniforme au-dessus de ce segment de droite. Ce point est la première observation dans

notre chaîne de Markov. Nous répétons alors le processus en remplaçant  $p_0$  par ce point pour obtenir un deuxième point aléatoire. En substituant ce deuxième point aléatoire à  $p_0$ , nous en obtenons un troisième, et ainsi de suite. De manière plus formelle, notre algorithme est :

Etape 1. Choisir un vecteur de probabilités positives initiales  $p_0$ , tel que  $A_{1,s} p_0 = b_1$  et  $A_{2,s} p_0 < b_2$  et fixer  $i = 0$ .

Etape 2. Produire une direction aléatoire  $d_i$  uniformément distribuée sur la sphère unitaire dans  $R^n$ .

Etape 3. Soit  $d_i^*$  la projection normalisée de  $d_i$  sur l'espace nul de  $A_{1,s}$ .

Etape 4. Trouver le segment de droite  $L_i = \{\alpha \in R | p_i + \alpha d_i^* \in P\}$  et générer  $\alpha_i$  uniformément sur ce segment de droite.

Etape 5. Fixer  $p_{i+1} = p_i + \alpha_i d_i^*$  et  $i = i + 1$  et retourner à l'étape 2.

À première vue, le rôle joué par les contraintes dans ce processus pourrait ne pas être clair. Cependant, elles interviennent par la voie de la définition de  $P$ . La chaîne de Markov produite de cette façon converge en loi vers la loi uniforme sur le polytope. Le résultat de convergence des algorithmes de mélange de ce type a été prouvé par Smith (1984). Si nous voulons approximer la valeur espérée d'une fonction définie sur  $P$ , la moyenne de la fonction calculée aux valeurs simulées converge vers sa valeur réelle. Cela permet de calculer des estimations ponctuelles des paramètres de population. Trouver approximativement l'intervalle de crédibilité bayésien au niveau à 0,95 est plus difficile.

Une possibilité consiste à exécuter la chaîne pendant une longue période; par exemple, nous pouvons générer 4,1 millions de valeurs, écarter les 100 000 premières et trouver les quantiles 0,025 et 0,975 des valeurs restantes. Ces deux nombres détermineront notre intervalle de crédibilité à 0,95 approximatif. Ici, nous considérons uniquement les tailles d'échantillon inférieures à 100. Dans ces conditions, nous avons constaté que des chaînes de quelques millions de valeurs de long suffisent.

La vitesse de mélange d'une chaîne dépend des contraintes et du paramètre estimé. Il semble que l'obtention d'un bon mélange prend plus de temps quand on estime la médiane que la moyenne, ce qui n'est pas étonnant si l'on se souvient que, dans les méthodes bootstrap standard, beaucoup plus d'échantillons bootstrap sont requis quand on

de population connue de  $x$ . Alors, étant donné un échantillon et les  $p_i$  satisfaisant  $\sum_{i \in S} p_i x_i = X$ , nous avons

$$E \left( \sum_{i \in S} p_i y_i \right) = aE \left( \sum_{i \in S} p_i \right) + bE \left( \sum_{i \in S} p_i x_i \right) + \sum_{i \in S} E(p_i e_i) \\ = a + bX \\ = \bar{y}_s - b\bar{x}_s + bX \\ = \bar{y}_s + b(\bar{X} - \bar{x}_s)$$

où  $b$  est l'estimation par les moindres carrés de  $b$ . Ici, les valeurs d'échantillon sont fixes, les  $p_i$  et les  $e_i$  sont aléatoires et l'espérance de  $p_i$  est calculée par rapport à la loi *a posteriori* de Pólya. La première approximation s'ensuit, puisque sous échantillonnage aléatoire simple, nous nous attendons à ce que les échantillons soient équilibrés, en moyenne, et que les  $p_i$  et les  $e_i$  soient approximativement indépendantes.

### 3.2 Contraintes linéaires et loi *a posteriori* de Pólya

L'information *a priori* concernant les variables auxiliaires peut avoir de multiples origines. Nous avons déjà discuté du cas où la moyenne de population d'une variable auxiliaire est connue. Un autre est celui où l'on connaît la médiane de population. De manière plus générale, on pourrait savoir uniquement qu'une moyenne ou une médiane de population est comprise dans un certain intervalle. Même si ce genre de cas est peu fréquentement discuté dans la littérature habituelle traitant de l'estimation fondée sur le plan de sondage, il semble assez raisonnable. Un autre cas est celui où une paire de variables auxiliaires décrit un tableau à double entrée dans lequel chaque unité doit appartenir à l'une des cellules et dont les totaux de population de ligne et de colonne pour les chiffres figurant dans chaque cellule sont connus. Avant de décrire l'approche de la loi *a posteriori* de Pólya contrainte pour résoudre ce genre de problème, nous devons mentionner un point technique mineur.

Supposons que nous sachions que la moyenne de la variable auxiliaire  $x$  est égale à  $\mu(x)$ . Il existera des échantillons pour lesquels la valeur de  $x$  est inférieure à  $\mu(x)$  pour chaque unité échantillonnée. Dans de tels cas, il sera impossible d'utiliser la loi *a posteriori* de Pólya contrainte. Toutefois, en pratique, cette situation ne se présente presque jamais. Nous supposons systématiquement que notre échantillon est « compatible » avec l'information *a priori*, ce que nous expliquons plus en détail plus loin. Dans nos études par simulation, nous rejetons systématiquement tout échantillon qui n'est pas compatible et nous en choisissons un autre. De nouveau, dans la plupart des cas, la probabilité de devoir rejeter un échantillon est très faible.

Chacun de nos exemples d'information *a priori* peut être représenté par une ou plusieurs contraintes linéaires d'égalité ou d'inégalité. Nous avons vu que connaître la moyenne de population donne une contrainte d'égalité linéaire. Si nous savons que la moyenne de population est comprise dans un certain intervalle, nous obtenons deux contraintes d'inégalité linéaires. Nous allons maintenant établir une notation qui nous permettra de considérer diverses situations où l'information *a priori* peut être décrite en utilisant des égalités et des inégalités linéaires comme contraintes.

Nous supposons qu'en plus de la caractéristique d'intérêt  $y$ , la population possède un ensemble de variables auxiliaires  $x_1, x_2, \dots, x_m$ . Pour l'unité  $i$ , soit

$$(y_i, x_1, x_2, \dots, x_m) = (y_i, x_1^i, x_2^i, \dots, x_m^i)$$

le vecteur de valeurs de  $y$  et de variables auxiliaires. Nous supposons que, pour toute unité échantillonnée, ce vecteur de valeurs est observé. Nous supposons aussi que l'information *a priori* au sujet de la population peut être exprimée à l'aide d'un ensemble de contraintes d'égalité et d'inégalité linéaires pondérées sur les valeurs auxiliaires distinctes dans la population, en prenant pour pondérations les proportions de la population qui prennent ces valeurs individuelles distinctes. Nous illustrons cet aspect plus en détail en expliquant comment nous transposons cette information *a priori* au sujet de la population aux valeurs observées dans l'échantillon, afin de pouvoir construire des pseudo-versions de la population compatibles avec l'information *a priori*.

Soit  $s$  un échantillon et, pour  $j = 1, 2, \dots, m(s)$ , soit  $(y_{ij}^s, x_{ij}^s)$  les valeurs observées que nous supposons distinctes pour simplifier. Soit  $d = (d^1, \dots, d^{n(s)})$  la proportion d'unités qui sont affectées à la valeur  $(y_{ij}^s, x_{ij}^s)$  dans la réplique complète simulée de la population. Toute contrainte linéaire sur les valeurs de population d'une variable auxiliaire se traduira d'une manière évidente par une contrainte linéaire sur les valeurs simulées. Par exemple, si nous savons que la moyenne de population de  $x^1$  est inférieure ou égale à une valeur donnée, disons  $b_1$ , pour la population simulée, cette information devient la contrainte

$$\sum_{i=1}^n d^1 x_{ij}^1 \leq b_1$$

Si nous savons que la médiane de population de  $x^2$  est égale à  $b_2$ , la contrainte pour la population simulée devient

$$\sum_{i=1}^n d^2 w_i = 0,5$$

où  $w_i = 1$  si  $x_{ij}^2 \leq b_2$  et est nul autrement. Si la moyenne de population de  $x^2$  est inférieure ou égale à la moyenne de

une extension de la loi *a posteriori* de Pólya (Meeden 1999).

La loi *a posteriori* de Pólya est semblable en esprit aux méthodes du bootstrap pour l'échantillonnage d'une population finie. Dans l'un et l'autre cas, un genre

d'argument d'échangeabilité est utilisé pour produire des pseudo-versions de la population. L'idée fondamentale du

bootstrap est décrite dans Gross (1980). Supposons que nous ayons un échantillon aléatoire simple de taille  $n(s)$

tiré de la population et supposons que  $N/n(s) = m$  est un

nombre entier. Partant de l'échantillon, nous créons une

bonne approximation de la population en combinant  $m$

répliques de l'échantillon. Puis, nous tirons des échantillons

aléatoires répétés de taille  $n(s)$  de cette population créée

pour étudier le comportement de l'estimateur d'intérêt. Les

propriétés asymptotiques des estimateurs peuvent également

être étudiées (voir Booth, Bulter et Hall 1994 pour des

précisions). Cette situation diffère de l'utilisation de la loi

*a posteriori* de Pólya qui, pour un échantillon fixe, produit

des versions complètes de la population et permet

d'examiner la distribution du paramètre d'intérêt dans la

population, plutôt que les propriétés de l'estimateur du

paramètre. Pour le paramètre de population d'intérêt, les

propriétés de l'estimateur découlent directement de cette loi

prédictive pour les valeurs de population.

La loi *a posteriori* de Pólya correspond au bootstrap

bayésien de Rubin (1981) appliqué à l'échantillonnage

d'une population finie. Le bootstrap bayésien original

s'applique à un échantillon aléatoire tiré d'une population

précédente.

Pour  $j = 1, \dots, n(s)$ , soit  $d_j$  la proportion d'unités dans

très utile quand nous examinerons la loi *a posteriori* de Pólya contrainte.

### 3. La loi *a posteriori* de Pólya contrainte

#### 3.1 L'idée fondamentale

Dans de nombreuses situations, en plus de la variable

d'intérêt,  $y$ , l'échantillonneur dispose de variables

auxiliaires,  $x$ , pour lesquelles existe de l'information

*a priori*. Par exemple, la moyenne de population,  $\mu_x$ , de  $x$

pourrait être connue. Étant donné une unité dans un

échantillon aléatoire, nous observons sa paire de valeurs

$(y, x)$ . En suivant notre notation antérieure, nous repré-

sentons l'échantillon par

$$z = (s, (y, x)_s) = (s, \{(y_i^s, x_i^s), \dots, (y_{n(s)}^s, x_{n(s)}^s)\}).$$

Dans cette situation, l'estimateur par la régression est

souvent utilisé pour estimer la moyenne de population.

Comment la loi *a posteriori* de Pólya devrait-elle être

ajustée pour tenir compte du fait que la moyenne de

population de  $x$  est connue? La réponse simple consiste à

restreindre la loi prédictive uniquement aux populations

compatibles avec l'information *a priori*. En pratique, cela

reviendrait à ne produire que des répliques complètes de la

population compatible avec l'information *a priori* connue.

Pour l'illustrer, nous considérons la forme approximative de

la loi *a posteriori* de Pólya décrite à la fin de la section

précédente.

Pour  $j = 1, \dots, n(s)$ , soit  $d_j$  la proportion d'unités dans

une réplique complète de la population qui ont la valeur

$(y_j, x_j)$ . Au lieu d'utiliser la loi uniforme pour  $d = (d_1, \dots, d_{n(s)})$  sur le simplexe pour générer les répliques

stimulées de la population, nous devrions utiliser la loi

uniforme limitée au sous-ensemble du simplexe satisfaisant

$$(1) \quad \sum_{n(s)} d_j x_j^i = \mu_x^i$$

Avant de décrire comment nous générerons les vecteurs de

$d$  à partir de la loi *a posteriori* de Pólya contrainte, nous

examinons le lien entre l'estimateur résultant et l'estimateur

par la régression.

De nombreux résultats de simulation (non présentés ici)

montrent que le comportement de la loi *a posteriori* de

Pólya contrainte ressemble fort à celui de l'estimateur par la

régression sous échantillonnage aléatoire simple. Le simple

argument qui suit montre pourquoi ces deux estimations

ponctuelles devraient concorder fréquemment, même si la

loi *a posteriori* de Pólya ne comporte aucune hypothèse au

sujet de la relation entre  $y$  et  $x$ .

Supposons une population  $y_i = a + b x_i + e_i$ , où  $e_i$  est

une erreur aléatoire d'espérance nulle. Soit  $\bar{X}$  la moyenne

des valeurs  $x_i$ , c'est-à-dire qui est uniforme sur le simplexe à

$n(s) - 1$  dimensions, où  $\sum_{n(s)} d_j = 1$ . Cette approche sera

Ditrichet avec un vecteur de paramètres qui ne contient que

$d = (d_1, \dots, d_{n(s)})$  suit approximativement une loi de

prend la valeur  $y_j$ . Alors, sous la loi *a posteriori* de Pólya,

Pour  $j = 1, \dots, n(s)$ , soit  $d_j$  la proportion d'unités dans

une réplique complète simulée de la population entière qui

Supposons que la fraction d'échantillonnage  $f$  soit faible.

approximation bien connue de la loi *a posteriori* de Pólya.

population, il est souvent plus efficace d'utiliser une

population finie. Dans l'un et l'autre cas, un genre

méthodes du bootstrap pour l'échantillonnage d'une

population finie. Dans l'un et l'autre cas, un genre

d'argument d'échangeabilité est utilisé pour produire des

pseudo-versions de la population. L'idée fondamentale du

bootstrap est décrite dans Gross (1980). Supposons que

nous ayons un échantillon aléatoire simple de taille  $n(s)$

tiré de la population et supposons que  $N/n(s) = m$  est un

nombre entier. Partant de l'échantillon, nous créons une

bonne approximation de la population en combinant  $m$

répliques de l'échantillon. Puis, nous tirons des échantillons

aléatoires répétés de taille  $n(s)$  de cette population créée

pour étudier le comportement de l'estimateur d'intérêt. Les

propriétés asymptotiques des estimateurs peuvent également

être étudiées (voir Booth, Bulter et Hall 1994 pour des

précisions). Cette situation diffère de l'utilisation de la loi

*a posteriori* de Pólya qui, pour un échantillon fixe, produit

des versions complètes de la population et permet

d'examiner la distribution du paramètre d'intérêt dans la

population, plutôt que les propriétés de l'estimateur du

paramètre. Pour le paramètre de population d'intérêt, les

propriétés de l'estimateur découlent directement de cette loi

prédictive pour les valeurs de population.

La loi *a posteriori* de Pólya correspond au bootstrap

bayésien de Rubin (1981) appliqué à l'échantillonnage

d'une population finie. Le bootstrap bayésien original

s'applique à un échantillon aléatoire tiré d'une population



Puisque, sous la loi *a posteriori* de Pólya, les seules valeurs de  $y$  qui figurent dans une réplique complète simulée de la population sont celles qui sont présentes dans l'échantillon, cette loi est simplement un moyen d'attribuer des poids aléatoires, c'est-à-dire des probabilités, aux unités comprises dans l'échantillon. Sous la loi *a posteriori* de Pólya, le poids moyen attribué à chaque unité échantillonnée est  $1/n(s)$ , de sorte que, comme nous l'avons vu, l'estimation de la moyenne de population soit simplement la moyenne d'échantillon. Ce sont cette relation et le caractère bayésien de la loi *a posteriori* de Pólya qui nous permettent de prouver l'admissibilité de la moyenne d'échantillon pour estimer la moyenne de population sous perte d'erreur quadratique. Cela donne à penser que les méthodes d'inférence fondées sur la loi *a posteriori* de Pólya auront tendance à concorder avec les méthodes fréquentistes et auront de bonnes propriétés fréquentistes.

À titre d'illustration supplémentaire de ce point, nous notons que récemment, deux des auteurs (Nelson et Meeden 2006) ont démontré que les intervalles de crédibilité bayésiens fondés sur la loi *a posteriori* de Pólya pour la médiane de population concordent asymptotiquement avec l'intervalle standard de Woodruff (Woodruff 1952). En guise d'autre exemple, considérons l'estimation de la moyenne ou du total d'une sous-population ou d'un domaine quand on utilise un échantillon aléatoire simple tiré de la population complète. Ici, le nombre d'unités échantillonnées qui appartiennent au domaine est une variable aléatoire. Donc, la moyenne des unités échantillonnées qui tombent dans le domaine est le ratio de deux variables aléatoires. Cette estimation est plus compliquée que celle de la moyenne de l'ensemble des unités dans l'échantillon. Pour obtenir une estimation de la variance pour cet estimateur, la méthode fréquentiste habituelle impose comme contrainte le nombre d'unités échantillonnées qui sont comprises dans le domaine. Cependant, quand on estime le total des domaines, cet argument conditionnel ne marche pas et il faut utiliser une méthode inconditionnelle pour obtenir une estimation de la variance. À cet égard, voir, par exemple, Cochran (1976). Récemment, l'un des auteurs (Meeden 2005) a montré que les inférences basées sur la loi *a posteriori* de Pólya concordent avec les réponses fréquentistes habituelles. Par conséquent, cette loi permet de traiter les deux situations en appliquant une seule théorie. Il est important de se rappeler que, dans l'approche fréquentiste, le conditionnement peut se faire sous l'échantillonnage aléatoire simple, mais que, pour des plans d'échantillonnage plus complexes, il n'est généralement pas faisable, puisque la loi aléatoire conditionnelle est inconnue. Enfin, notons que les méthodes d'échantillonnage par grappe à deux degrés fréquentistes habituelles peuvent être justifiées pour

Cette loi prédictive produit souvent des estimateurs semblables aux estimateurs fréquentistes classiques sous l'estimation de la moyenne de population. Avant de poursuivre, nous devons étendre un peu sur la notation. Soit  $s$  un échantillon possible de taille  $n(s)$ . Il s'agit d'un sous-ensemble de  $\{1, 2, \dots, N\}$ , l'ensemble d'éléments pour la population finie. Si  $s = \{i_1, \dots, i_{n(s)}\}$ , alors  $y_s = \{y_{i_1}, \dots, y_{i_{n(s)}}\}$  est l'ensemble de valeurs observées pour  $y$ , la caractéristique d'intérêt. Soit  $z = (s, y_s)$  un échantillon observé typique. Alors, sachant  $z = (s, y_s)$ , nous avons

$$\bar{z}_s = \sum_{i \in s} y_i / n(s)$$

$$\text{Var}(z) = \sum_{i \in s} (y_i - \bar{z}_s)^2 / (n(s) - 1)$$

qui sont la moyenne et la variance d'échantillon. Soit  $y^{(m)}(y) = \sum_{i=1}^m y_i / m$  la moyenne de population. Sous la loi *a posteriori* de Pólya,

$$E(y^{(m)}(y) | z) = \bar{z}_s$$

$$\text{Var}(y^{(m)}(y) | z) = (1 - f) \frac{m(s)}{n(s)} - 1$$

où  $f = n(s)/N$ . Notons que, sauf pour le dernier facteur de la variance *a posteriori*, ces deux termes sont simplement la moyenne d'échantillon et sa variance sous échantillonnage aléatoire simple. Les probabilités données par le plan de sondage ne jouent aucun rôle explicite dans ces calculs. Néanmoins, pour que la loi *a posteriori* de Pólya soit jugée appropriée par le spécialiste de l'échantillonnage, il faut que les valeurs de la caractéristique d'intérêt pour les unités observées et non observées soient approximativement interchangeables. C'est dans ce genre de situation que l'échantillonnage aléatoire simple sans remise est utilisé.

Sous la loi *a posteriori* de Pólya, il n'est pas toujours possible de trouver explicitement l'intervalle bayésien de crédibilité pour la moyenne de population ou pour les estimations ponctuelles et par intervalle d'autres paramètres de population. Dans de tels cas, il est facile de trouver ces estimations approximativement en simulant répétitivement des répliques complètes de la population. Pour chaque réplique simulée, nous calculons le paramètre de population d'intérêt. L'expérience a montré que de 500 à 1 000 valeurs simulées donneront habituellement de bons résultats. La moyenne de ces valeurs calculées sera notre estimation ponctuelle et les quantiles 0,025 et 0,975 de ces valeurs calculées donneront notre estimation par intervalle.





Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires

Radu Lazar, Glen Meeden et David Nelson

## Résumé

L'enchantillonnage de populations finies, nous ne dispose souvent d'information *a priori* sous la forme de renseignements partiels sur une variable auxiliaire, dont on connaît, par exemple, la moyenne. Dans de tels cas, on utilise fréquemment l'estimateur par le ratio et l'estimateur par la régression pour estimer la moyenne de population de la caractéristique d'intérêt. La loi *a posteriori* de Polya a été établie à titre d'information bayésienne non informative de l'enchantillonnage. Ici, il n'existe aucune, au sujet de la population, d'information *a priori*, voire aucune, au sujet de la population, qui peut être étendue afin d'y intégrer divers types d'information *a priori* partielle au sujet des variables auxiliaires. Nous verrons que cette approche produit généralement des méthodes ayant de bonnes propriétés fréquentistes, même dans certains cas où les méthodes fréquentistes classiques sont difficiles à appliquer.

Mots clés : Echantillonnage; approche bayésienne non informative; variable auxiliaire; contraintes linéaires et loi *a posteriori* de Polya.

## 1. Introduction

Charmatz (1986), Kassar, Kassar et Mantel (1990), ainsi que  
Charmatz (1986), Chénier (1993), Chénier (1999), Kassar,  
Kassar et Mantel (1990), ainsi que Charmatz (1986).

Chambers et Dunstan (1986). L'un des avantages de l'approche bayésienne est qu'une loi *a posteriori* raisonnable pour la population permet d'intégrer l'information *a priori* dans l'estimation de plusieurs paramètres de population. Même si la loi *a posteriori* n'a pas d'expression explicite pour un estimateur particulier, pour tout échantillon donné, nous pouvons trouver sa valeur approximative. Pour cela, nous échantillonons la loi *a posteriori* pour simuler des répliques complètes de la population et nous employons des méthodes d'estimation de Monte Carlo. Si la loi *a posteriori* n'a pas de forme commode pour l'échantillonnage, il devrait être possible d'utiliser des méthodes Monte Carlo par chaînes de Markov pour mettre en œuvre le processus de simulation. Pour chacune des répliques simulées, nous calculons la valeur du paramètre d'intérêt. En simulant un grand nombre de répliques complètes de la population, nous pouvons trouver, approximativement, les estimations bayésiennes ponctuelles et par intervalle correspondantes du paramètre de population d'intérêt. Le problème se résume alors à trouver un modèle bayésien de population raisonnable qui utilise le genre d'information *a priori* disponible pour la variable auxiliaire. Souvent, on peut fonder des modèles bayésiens raisonnables sur la loi *a posteriori* de Polya. Cette dernière représente une méthode bayésienne non informative d'échantillonnage d'une population finie qui n'utilise que peu d'information *a priori*, voire aucune au sujet de

L'échantillonnage de populations finies est l'un des domaines de la statistique où l'information *a priori* est utilisée systématiquement pour faire des inférences. Dans la plupart des cas, cette information *a priori* n'est pas intégrée dans le problème de manière bayésienne, entre autres raisons, parce qu'elle ne mène habituellement pas, d'une manière évidente, à une loi *a priori* raisonnable. Dans l'approche prédictive (voir Vallian, Dorfman et Royall, 2000), on part d'un modèle hypothétique et l'on estime ses paramètres inconnus après avoir observé l'échantillon. Dans la théorie fréquentiste classique, l'information *a priori* est encapsulée dans le plan d'échantillonnage probabiliste ou plan de sondage. Basu a montré que, pour un tenant de la théorie bayésienne, une fois que l'échantillon a été observé, le plan d'échantillonnage ne joue plus aucun rôle dans la loi *a posteriori*. [Pour ce fait et d'autres réflexions de Basu sur l'échantillonnage de populations finies, voir Ghosh (1988).] Bien que cette conclusion radicale ne soit pas acceptée uniformément par les bayésiens, il est vrai que la théorie fréquentiste classique et l'approche bayésienne de l'échantillonnage ont peu de chose en commun.

La théorie classique de l'échantillonnage met l'accent sur l'estimation de la moyenne de population. Si celle-ci est connue *a priori* pour une variable auxiliaire, on utilise souvent l'estimateur par le ratio ou l'estimateur par la régression. Si l'on souhaite estimer un autre paramètre que la moyenne, disons un quantile de population ou la fonction de répartition de la population, ou que l'on possède de l'information *a priori* sur une autre variable auxiliaire que

1. Radu Lazar, School of Statistics, University of Minnesota, MN 55455. Contact: [radu@stat.umn.edu](mailto:radu@stat.umn.edu); Glen Meeden, Center for Chronic Disease Outcomes Research, Minneapolis VA Medical Center, Minneapolis, MN 55417. Contact: [Dave.Nelson@med.vad.gov](mailto:Dave.Nelson@med.vad.gov).



Remerciements

Les présents travaux ont été réalisés pendant que Balgobin Nandram était en congé sabbatique au National Center for Health Statistics, à Hyattsville, au Maryland, en 2003-2004.

Annexe A

Modèle variable en fonction du temps : densités a posteriori conditionnelles de  $p_t, t = 0, \dots, T$

Nous montrons comment tirer un échantillon de la densité a posteriori conditionnelle de  $p_0$  dans (17) en utilisant une méthode par grille, et comment tirer un échantillon des densités a posteriori conditionnelles des  $p_t, t = 1, \dots, T$  dans (18), en utilisant des pas de Metropolis, chacun avec une chaîne indépendante.

En premier lieu, nous montrons comment tirer un échantillon d'une densité a posteriori conditionnelle de  $p_0$  dans (17) par une méthode de grille. En posant que  $(q_{01}, \dots, q_{0L}) = (p_{01}, \dots, p_{0L})$  et  $(q_{11}, \dots, q_{1L}) = (p_{11}, \dots, p_{1L})$  où  $L = r_C$ , avec  $\sum_{\ell=1}^{L-1} q_{0\ell} \leq 1$ , nous avons

$$\pi(q_{01}, \dots, q_{0L-1} | \text{sinon}, y_1) \\ \propto \frac{\prod_{\ell=1}^{L-1} q_{0\ell}^{q_{0\ell}} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{1 - \sum_{\ell=1}^{L-1} q_{0\ell}}}{\prod_{\ell=1}^{L-1} q_{0\ell}^{q_{0\ell}} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{1 - \sum_{\ell=1}^{L-1} q_{0\ell}}}$$

et il est facile de montrer que

$$\pi(q_{0\ell} | \text{sinon}, y_1) \propto \frac{\prod_{\ell=1}^{L-1} q_{0\ell}^{q_{0\ell}} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{1 - \sum_{\ell=1}^{L-1} q_{0\ell}}}{\prod_{\ell=1}^{L-1} q_{0\ell}^{q_{0\ell}} (1 - \sum_{\ell=1}^{L-1} q_{0\ell})^{1 - \sum_{\ell=1}^{L-1} q_{0\ell}}}$$

$$0 \leq q_{0\ell} \leq 1 - \sum_{\ell=1}^{L-1} q_{0\ell}, \ell = 1, \dots, L - 1.$$

Pour chaque  $\ell$ , nous divisons l'intervalle  $0 \leq q_{0\ell} \leq 1 - \sum_{\ell=1}^{L-1} q_{0\ell}$  en plusieurs sous-intervalles. Pour tirer une valeur aléatoire  $q_{0\ell}$  de sa densité a posteriori conditionnelle, nous sélectionnons un intervalle proportionnel à son aire et tirons une valeur aléatoire uniforme de cet intervalle. En deuxième lieu, nous montrons comment tirer un échantillon des densités a posteriori conditionnelles des  $p_t, t = 1, \dots, T$  dans (18), en utilisant des pas de Metropolis, chacun avec une chaîne indépendante. Considérons  $p_t | p_{-t}, \tau_2, y, t = 1, \dots, T$ . Nous utilisons la densité générant les valeurs candidates

$$p_t | p_{-t}, \tau_2, y \sim \text{Dirichlet}(a_t).$$

$$\text{où } a_{jk} = y_{j-1,k} + \tau_2 p_{j-1,k}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c.$$

Bibliographie

Chang, L.C., et Krosnick, J.A. (2001). Improving election forecasting. *Technical Report*, Department of Psychology, The Ohio State University.

Chen, Q.L., et Stasny, E.A. (2003). Handling undecided voters: Using missing data methods in election forecasting. *Technical Report*, Department of Statistics, The Ohio State University.

Chen, T., et Fienberg, S.E. (1974). Two-dimensional contingency tables with both complete and partially cross-classified data. *Biometrics*, 30, 629-642.

DeSart, J., et Holbrook, T.M. (2003). Campaigns, polls, and the states: Assessing the accuracy of statewide presidential trial-heat polls. *Political Research Quarterly*, 56, 431-439.

Draper, D. (1995). Assessment and propagation of model uncertainty (avec discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45-97.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Gelman, A., Roberts, G.O., et Gilks, W.R. (1996). Efficient Metropolis jumping rules. Dans *Bayesian Statistics*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid et A.F.M. Smith), Oxford, U.K.: Oxford University Press, 599-607.

Ghosh, M., et Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. New York : Chapman & Hall.

Kalton, G., et Kasprzyk, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

Lavrakas, P.J. (1993). *Telephone Survey Method: Sampling, Selection, and Supervision*. Newbury Park, CA : Sage Publications.

Little, R.J.A., et Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2<sup>me</sup> édition, New York : John Wiley & Sons Inc.

Nandram, B. (1998). A Bayesian analysis of the three-stage hierarchical multinomial model. *Journal of Statistical Computation and Simulation*, 61, 97-126.

Nandram, B., et Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., et Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., Cox, L.H., et Choi, J.W. (2005). Analyse bayésienne des données catégoriques manquantes non ignorables : une application à la densité minérale osseuse et au revenu familial. *Techniques d'enquête*, 31, 233-247.

Ritter, C., et Tanner, M.A. (1992). The Gibbs stopper and the Griddy Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association*, 73, 40-46.

Alors, la probabilité d'acceptation est  $\min(1, \psi_{s+1}/\psi_s)$ , où 
$$\psi_s = \prod_{r=1}^c \prod_{k=1}^{p_{s+1,k}^{(s)}} p_{t+1,k}^{(s)} / D(p_{(s)}^{(s)}; \tau_2).$$

Tableau 4  
Sensibilité des moyennes *a posteriori* (MP) et des écarts-types *a posteriori* (ETP) de  $\Delta_p$  et  $\Delta_p$  en fonction des variations de  $k_1$  et  $k_2$  par modèle

Modèle	$k_1$			1			5			25			50					
	$\Delta_p$	Non ignorable	Variable en fonction du temps	$\Delta_p$	Non ignorable	Variable en fonction du temps	$\Delta_p$	Non ignorable	Variable en fonction du temps	$\Delta_p$	Non ignorable	Variable en fonction du temps	$\Delta_p$	Non ignorable	Variable en fonction du temps			
NOTA : Nous avons pris $\tau_1 \sim \text{Gamma}(1/k_1^2, 1/k_1^2)$ et nous avons étudié la sensibilité en fonction de $k_1$ et de $k_2$ . Voir la note du tableau 3.	1	0,058	0,150	0,046	0,153	0,060	0,148	0,054	0,147	1	0,663	0,200	0,650	0,194	0,666	0,186	0,670	0,182
	2	0,051	0,153	0,046	0,146	0,062	0,151	0,054	0,145	2	0,663	0,197	0,661	0,188	0,667	0,185	0,659	0,181
	3	0,058	0,152	0,059	0,145	0,053	0,149	0,055	0,149	3	0,663	0,199	0,647	0,196	0,666	0,184	0,669	0,180
	4	0,055	0,151	0,057	0,148	0,049	0,148	0,043	0,154	4	0,641	0,202	0,668	0,191	0,643	0,197	0,650	0,195
	1	0,050	0,134	0,044	0,144	0,048	0,136	0,050	0,130	1	0,648	0,155	0,642	0,123	0,657	0,099	0,661	0,095
	2	0,049	0,136	0,052	0,140	0,056	0,129	0,047	0,137	2	0,660	0,151	0,652	0,127	0,659	0,102	0,657	0,099
	3	0,039	0,139	0,049	0,137	0,045	0,139	0,052	0,133	3	0,622	0,153	0,636	0,137	0,649	0,120	0,648	0,115
	4	0,037	0,138	0,042	0,138	0,041	0,141	0,051	0,129	4	0,610	0,162	0,636	0,152	0,646	0,132	0,644	0,127

Notre modèle variable en fonction du temps fournit des inférences *a posteriori* plus proches des valeurs réelles que les modèles de non-réponse ignorable et non ignorable, et que ceux de Chen et Stasny (2003). Il est naturel que les préférences des électeurs évoluent à mesure que de nouveaux renseignements, nuisibles ou favorables, sont révélés au public. Donc, notre modèle de non-réponse variable en fonction du temps, qui tient compte des changements au cours du temps et donne une meilleure précision, est celui auquel il faut donner la préférence. L'incertitude de la prédiction peut être réduite de deux manières. Premièrement, l'accroissement du nombre de paramètres, ce qui à son tour pourra améliorer la prédiction. Deuxièmement, une plus grande quantité d'information *a priori* (par exemple, sondage à la sortie de l'isoloir) au sujet de l'élection de novembre permettra également d'améliorer la prédiction.

Nos intervalles de crédibilité à 95 % peuvent être raccourcis en utilisant des renseignements *a priori* sur la proportion d'électeurs votant pour Taft ou pour Fisher. Selon un examinateur, « dans les élections générales, les proportions de votes pour les grands partis sont comprises entre 35 % et 65 % et pour des états particuliers, un politologue objectif pourrait généralement fournir un *prior* encore plus serré ». Cependant, le problème est complexe, parce que, quand les distributions *a priori* sont trouvées sur les  $p$ , il existe une constante de normalisation qui est une fonction de  $\tau_2$ . Donc, quand  $\tau_2$  est tiré de la densité

Le nombre de jours avant un élection a un effet important sur l'exactitude des sondages et cet effet peut varier considérablement selon le contexte de la campagne électorale (voir, par exemple, Desart et Holbrook 2003). Il est donc fort difficile de prédire le résultat d'une élection plusieurs semaines avant qu'elle ait effectivement lieu, à moins qu'il existe une marge absolue. Une personne qui souhaite prédire le résultat d'une élection doit tenir compte de l'information supplémentaire aux approches de l'élection réelle. Notre prédiction est fondée sur l'hypothèse qu'aucun changement catastrophique ne surviendra à l'approche de l'élection; mais ce genre de changement soudain dans l'opinion publique peut avoir lieu. Par exemple, en 1988, Dukakis a perdu l'élection contre George Bush pour diverses raisons, dont le fait qu'il a passé la dernière semaine de la campagne au Massachusetts, que sa personnalité était jugée froide et que Bush a attaqué sa position libérale. En outre, une campagne efficace peut mobiliser les électeurs indécis à l'approche de l'élection (par exemple, Truman et Dewey en 1948). Un moyen de tenir compte d'une catastrophe éventuelle consiste à utiliser un mélange de distribution ou d'autres distribution à queues lourdes (à l'instar des chercheurs qui utilisent des distributions de Levy en mathématiques financières).

$\Delta_e = \Pr(q_2 > q_1 | \alpha)$

$$= \int_{1/2}^0 \int_{-q_1}^{-q_2} \frac{q_1}{q_2} \frac{q_2}{q_1} \frac{D(\alpha_1, \alpha_2, \alpha_3)}{D(\alpha_1, \alpha_2, \alpha_3)} dq_2 \Bigg\} dq_1.$$

Il est alors facile de montrer que

$$\Delta_e = 1 - F_{\alpha_1, \alpha_2 + \alpha_3}(1/2) \int_{1/2}^0 F_{\alpha_2, \alpha_3}\{q_1/(1 - q_1)\}$$

$$\Bigg\} \frac{F_{\alpha_1, \alpha_2 + \alpha_3}}{q_1^{\alpha_1 - 1}(1 - q_1)^{\alpha_2 + \alpha_3 - 1}} \Bigg\} dq_1, \quad (26)$$

ou

$$F_{\alpha_1, \alpha_2 + \alpha_3}(a) = \int_a^0 \frac{x^{\alpha_1 - 1}(1 - x)^{\alpha_2 + \alpha_3 - 1}}{B(\alpha_1, \alpha_2 + \alpha_3)} dx$$

et

$$F_{\alpha_2, \alpha_3}(a) = \int_a^0 \frac{x^{\alpha_2 - 1}(1 - x)^{\alpha_3 - 1}}{B(\alpha_2, \alpha_3)} dx.$$

Nous notons que  $\Delta_e$  est la probabilité que Taft obtienne une plus forte proportion de votes au sondage d'octobre et que  $\Delta_p$  est la probabilité que Taft reçoive une plus forte proportion de votes à l'élection de novembre. Ces paramètres peuvent être fort utiles pour l'estimation ( $e$ ) et la prédiction ( $p$ ). Des paramètres tels que  $\Delta_e$  et  $\Delta_p$  sont difficiles à analyser selon une approche non bayésienne telle que celle de Chen et Stasny (2003), il s'agit effectivement d'une grande force du paradigme bayésien.

Il est facile de calculer (26) par une intégration de Monte Carlo. Pour chaque  $\alpha_1, \alpha_2, \alpha_3, \Delta_e q_1 \sim \text{Béta}(\alpha_1, \alpha_2 + \alpha_3)$  tronquée à l'intervalle  $(0, 1/2)$  est utilisée comme une fonction d'importance. Par conséquent, pour chaque  $\alpha_{(h)}, h = 1, \dots, M, M = 1\,000$  tiré de l'échantillonneur de Metropolis-Hastings (ou de l'échantillonneur de Gibbs), nous pouvons calculer  $\Delta_{(h)}^e$ . Nous obtenons l'inférence  $a posteriori$  au sujet de  $\Delta_e$  de la manière empirique classique. Pour la prédiction, nous avons considéré  $\Delta_p = \Pr(p_{412} > p_{411} | p_{412} + p_{413}, \alpha)$ , où  $\alpha_{jk} = \mu_{jk} \tau_{2j}$ ,  $j = 1, \dots, r, k = 1, \dots, c$ , pour les modèles de non-réponse ignorables et non ignorables, et  $\alpha_{jk} = p_{3jk} \tau_{2j}$ ,  $j = 1, \dots, r, k = 1, \dots, c$ , pour le modèle de non-réponse variable en fonction du temps. Notons que  $\Delta_e$  et  $\Delta_p$  sont les mêmes pour les modèles de non-réponse ignorables et non ignorables.

Au tableau 3, nous présentons aussi les estimations de  $\Delta_e$  et  $\Delta_p$  pour les trois modèles. Premièrement, soulignons de nouveau que toutes les ETN sont faibles. Les estimations de ces paramètres sont semblables pour les trois modèles et supérieures à 0,60, mais les intervalles de crédibilité à 95 % contiennent la valeur 0,5. Par conséquent, de nouveau, les moyennes  $a posteriori$  indiquent que Taft est le gagnant, mais la variation n'illustre l'effet de Taft étant le gagnant. Nous remarquons une fois de plus que le modèle de non-réponse variable en fonction du temps fournit une inférence plus précise, quoique pas suffisamment. Les

### Conclusion

La principale contribution du présent article comprend la construction et l'analyse d'un modèle de non-réponse non ignorable variable en fonction du temps et son application aux données des sondages électoraux en Ohio. À cela s'ajoutent deux autres accomplissements. Premièrement, nous avons comparé le modèle variable en fonction du temps à une version étendue (afin d'y inclure la variable de temps) des modèles de non-réponse ignorables et non ignora- bles de Nandram et coll. (2005). Deuxièmement, nous avons construit un nouveau paramètre pour faciliter la prédiction du gagnant; cependant, ce paramètre n'a pas changé énormément les choses, en partie parce que le modèle de non-réponse variable en fonction du temps ne comporte que trois points dans le temps.

Notons qu' $a priori$

$$\tau_1 \sim \text{Gamma}\left(\frac{1}{k_1^2 c_0^2}, \frac{1}{k_2 \mu_0 k_1^2 c_0^2}\right),$$

$E(\tau_1) = k_2 \mu_0$  et  $E^{-1}(\tau_1) = k_1 k_2 c_0 \mu_0$ ; par conséquent, manifestement,  $E(\tau_1)$   $a priori$  augmente avec  $k_2$  et  $E^{-1}(\tau_1)$  augmente avec  $k_1$  ou  $k_2$ , mais pas nécessairement  $a posteriori$ . Ces changements n'ont pas beaucoup d'effet sur l'inférence  $a posteriori$ . Pour presque toutes les combinaisons de  $k_1$  et  $k_2$ , sous le modèle de non-réponse variable en fonction du temps, les écarts-types  $a posteriori$  de  $\Delta_p$  sont plus faibles (mais pas considérablement) que sous le modèle de non-réponse non ignorable. Sous le modèle de non-réponse variable en fonction du temps, les écarts-types  $a posteriori$  de  $\Delta_p$  sont plus faibles (mais pas considérablement plus petits que sous le modèle de non-réponse non ignorable pour toutes les combinaisons de  $k_1$  et  $k_2$ .

Enfin, nous étudions la sensibilité à l'inférence au sujet vue probabiliste, ils sont plus appropriés.

limitent l'inférence à une région plus petite en conditionnant sur  $p_{311} + p_{312} + p_{313}$  et  $p_{411} + p_{412} + p_{413}$ , et d'un point de paramètres  $\Delta_e$  et  $\Delta_p$  sont plus sensibles, parce qu'ils



le gagnant, la variabilité ne révèle aucune différence entre Taft et Fisher. Nous reviendrons sur ce point plus loin.

**Tableau 2**  
Comparaison de la proportion d'électeurs probables pour le sondage d'octobre 1998 et de la prédiction pour l'élection de novembre 1998 sous divers modèles avec le résultat réel

Intention	Fisher	Taft	Autre
Estimation d'échantillon	0,41	0,51	0,08
IC à 95 % d'approximatif	(0,35; 0,47)	(0,45; 0,57)	(0,05; 0,11)
Résultat réel	0,45	0,50	0,05

a. Estimation

Modèles A, B, C de Chen/Stasny	0,41	0,51	0,08
Modèle D de Chen/Stasny	0,42	0,51	0,07
Modèle E de Chen/Stasny	0,41	0,51	0,08
Modèle de non-réponse ignorable	0,41	0,51	0,08
IC à 95 %	(0,35; 0,46)	(0,46; 0,57)	(0,05; 0,12)
Modèle de non-réponse non ignorable	0,41	0,50	0,09
IC à 95 %	(0,32; 0,51)	(0,40; 0,60)	(0,05; 0,17)
Modèle variable en fonction du temps	0,41	0,50	0,09
IC à 95 %	(0,32; 0,52)	(0,41; 0,60)	(0,05; 0,16)

b. Prédiction

Modèle de non-réponse ignorable	0,41	0,54	0,05
IC à 95 %	(0,15; 0,70)	(0,25; 0,81)	(0,00; 0,22)
Modèle de non-réponse non ignorable	0,42	0,52	0,06
IC à 95 %	(0,15; 0,70)	(0,22; 0,79)	(0,00; 0,28)
Modèle variable en fonction du temps	0,41	0,50	0,09
IC à 95 %	(0,15; 0,71)	(0,21; 0,78)	(0,00; 0,31)

NOTA :  $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2)$ , où  $c_0 = 0,031$  et  $\mu_0 = 2,431$ .

Bien que nos estimations à l'aide du modèle de non-réponse variable en fonction du temps soient proches des estimations réelles, les intervalles de crédibilité à 95 % pour  $p_{311}$  et  $p_{312}$  se chevauchent, si bien qu'il est difficile de prédire que Taft est le gagnant. Les intervalles de crédibilité à 95 % sont plus courts pour nos autres modèles, mais les estimations ponctuelles ne sont pas aussi bonnes et ces intervalles se chevauchent encore. L'une des faiblesses de notre analyse au tableau 2 tient au fait que nous n'avons pas tenu compte de la corrélation entre les deux estimations (autrement dit, nous devrions vraiment étudier la différence  $p_{312} - p_{311}$ , qui est la marge pour gagner).

Au tableau 3, nous présentons les estimations de  $\Delta_e = p_{312} - p_{311}$  et  $\Delta_p = p_{412} - p_{411}$  à  $k_1 = k_2 = 1$  pour les trois modèles. Nous avons également inclus l'erreur-type numérique (ETN), qui est une mesure du degré de reproductibilité des résultats numériques; nous l'avons calculée par la méthode des moyennes par lot. De faibles ETN signifient que, si nous répétions le calcul complet de la même façon (c'est-à-dire en utilisant une autre série de 1 000 itérations), nous ne devrions observer que peu de différence entre les deux ensembles de réponse. Au tableau 3, les ETN sont petites. Les estimateurs et les

prédicteurs ponctuels sont tous positifs, indiquant que Taft est le gagnant au sondage d'octobre ainsi qu'à l'élection de novembre. Cependant, la variabilité éclipse dans une certaine mesure ce résultat, parce que les écarts-types *a posteriori* (ETP) sont grands, et, comme prévu, encore plus dans le cas des prédictions. Par conséquent, pour les deux paramètres, les intervalles de crédibilité à 95 % contiennent la valeur 0. Donc, de nouveau, si l'on considère la variabilité, aucune différence ne se dégage entre Taft et Fisher.

**Tableau 3**  
Comparaison des trois modèles pour l'estimation et la prédiction en utilisant les moyennes *a posteriori* (MP), les écarts-types *a posteriori* (ETP), les erreurs-types numériques (ETN) et les intervalles de crédibilité à 95 % pour  $\Delta_e$  ( $\Delta_p$ ) et  $\Delta_e$  ( $\Delta_p$ )

Modèle	MP	ETP	ETN	Intervalle
$\Delta_e$				
Ignorable	0,105	0,055	0,002	(-0,002; 0,209)
Non ignorable	0,097	0,099	0,006	(-0,100; 0,280)
Variable en fonction du temps	0,093	0,101	0,007	(-0,098; 0,276)
Ignorable	0,071	0,154	0,004	(-0,240; 0,362)
Non ignorable	0,058	0,150	0,005	(-0,252; 0,369)
Variable en fonction du temps	0,050	0,134	0,005	(-0,244; 0,314)
Ignorable	0,688	0,175	0,008	(0,295; 0,958)
Non ignorable	0,663	0,200	0,012	(0,222; 0,959)
Variable en fonction du temps	0,632	0,148	0,014	(0,336; 0,901)
Ignorable	0,688	0,175	0,008	(0,295; 0,960)
Non ignorable	0,663	0,193	0,009	(0,253; 0,972)
Variable en fonction du temps	0,648	0,155	0,011	(0,341; 0,923)

NOTA : Voir la note du tableau 2;  $\Delta_e = p_{312} - p_{311}$  (estimation, d'octobre);  $\Delta_p = p_{412} - p_{411}$  (prédiction, différence entre Taft et Fisher pour l'élection de novembre);  $\Delta_e = \text{Pr}(p_{312} > p_{311} | p_{311} + p_{312} + p_{413}; \alpha_e)$ ; et  $\Delta_p = \text{Pr}(p_{412} > p_{411} + p_{412} + p_{413}; \alpha_p)$ ; voir (26).

Nous cherchons un paramètre de rechange qui nous aidera à prédire le gagnant de manière plus convaincante. Nous posons la question suivante : « Quelle est la probabilité que la proportion de personnes votant pour Taft au moment du sondage d'octobre et au moment de l'élection de novembre soit plus grande que celle de personnes votant pour Fisher? »

Donc, nous considérons le paramètre  $\Delta_e = \text{Pr}(p_{312} > p_{311} | p_{311} + p_{312} + p_{413}; \alpha_e)$ , où  $\alpha_{jk} = \mu_{jk} \tau_{2j} = j, \dots, r, k = 1, \dots, c_j$  pour les modèles de non-réponse ignorable et non ignorable, et  $\alpha_{jk} = p_{2jk} \tau_{2j} = j, \dots, r, k = j, \dots, c_j$  pour le modèle de non-réponse variable en fonction du temps. Dans chaque cas, en posant  $q_1 = p_{311}/p_{312}$ ,  $q_2 = p_{312}/p_{311}$ , et  $q_3 = p_{313}/p_{311}$ , avec  $p_{311} = \sum_{k=1}^r p_{31k}$  et  $\sum_{k=1}^r q_k = 1$ , il est facile de montrer que  $(q_1, q_2, q_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$ , où  $\alpha_1 = \alpha_{11}$ ,  $\alpha_2 = \alpha_{12}$  et  $\alpha_3 = \alpha_{13} + \sum_{k=1}^r \alpha_{2k}$ . Par conséquent, nous avons

Pour étudier la sensibilité à l'erreur de spécification de la densité *a priori* de  $\tau_1$ , nous utilisons deux constantes,  $\kappa_1$  et  $\kappa_2$ , telles que, *a priori*

$$\tau_1 \sim \text{Gamma}(1/\kappa_1^2, c_0^2) \quad 1/\kappa_1^2 \leq \kappa_2 \mu_0 c_0^2$$

pour des valeurs variables de  $\kappa_1$  et  $\kappa_2$ . Il convient de souligner que  $E(\tau_1) = \kappa_2 \mu_0$  et  $\text{CV}(\tau_1) = \kappa_1 c_0$ ; donc, accroître la valeur de  $\kappa_2$  signifie accroître celle de  $\tau_1$ , ce qui, à son tour, signifie accroître la précision *a priori*, mais pas nécessairement *a posteriori*. Nous étudierons la sensibilité à la spécification de  $\kappa_1$  et de  $\kappa_2$  quand nous décrirons l'analyse des données.

### 3.5 Estimation et prédiction

Nous montrons comment améliorer l'estimation (c'est-à-dire, *rao-blackwellisation*) pour le sondage d'octobre et comment calculer la prédiction pour l'élection de novembre. Pour les modèles de non-réponse ignorable et non ignorable,

$$g(\mathbf{p}_T | \mathbf{y}_1) = \int g(\mathbf{p}_T | \boldsymbol{\mu}_2, \tau_2) \pi(\boldsymbol{\mu}_2, \tau_2 | \mathbf{y}_1) d\boldsymbol{\mu}_2 d\tau_2 \quad (22)$$

$$\approx \frac{1}{M} \sum_{h=1}^M g(\mathbf{p}_T | \boldsymbol{\mu}_{(h)}^*, \tau_{(h)}^*),$$

où  $\mathbf{p}_T | \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}(\boldsymbol{\mu}_2, \tau_2)$ , et, pour le modèle variable en fonction du temps,

$$g(\mathbf{p}_T | \mathbf{y}_1) = \int g(\mathbf{p}_T | \mathbf{p}_{T-1}, \tau_2) \pi(\mathbf{p}_{T-1}, \tau_2 | \mathbf{y}_1) d\mathbf{p}_T d\tau_2 \quad (23)$$

$$\approx \frac{1}{M} \sum_{h=1}^M g(\mathbf{p}_T | \mathbf{p}_{(h)}^{T-1}, \tau_{(h)}^*),$$

où  $\mathbf{p}_T | \mathbf{p}_{T-1}, \tau_2 \sim \text{Dirichlet}(\mathbf{p}_{T-1}, \tau_2)$ . Nous obtenons (prédisons) les proportions de cellule pour novembre comme il suit. Pour le modèle de non-réponse ignorable ou non ignorable, la densité *a posteriori* de  $\mathbf{p}_{T+1}$  est

$$g(\mathbf{p}_{T+1} | \mathbf{y}_1) = \int g(\mathbf{p}_{T+1} | \boldsymbol{\mu}_2, \tau_2) \pi(\boldsymbol{\mu}_2, \tau_2 | \mathbf{y}_1) d\boldsymbol{\mu}_2 d\tau_2 \quad (24)$$

$$\approx \frac{1}{M} \sum_{h=1}^M g(\mathbf{p}_{T+1} | \boldsymbol{\mu}_{(h)}^*, \tau_{(h)}^*),$$

où  $\mathbf{p}_{T+1} | \boldsymbol{\mu}_2, \tau_2 \sim \text{Dirichlet}(\boldsymbol{\mu}_2, \tau_2)$ . Pour le modèle variable en fonction du temps,

$$g(\mathbf{p}_{T+1} | \mathbf{y}_1) = \int g(\mathbf{p}_{T+1} | \mathbf{p}_T, \tau_2) \pi(\mathbf{p}_T, \tau_2 | \mathbf{y}_1) d\mathbf{p}_T d\tau_2 \quad (25)$$

$$\approx \frac{1}{M} \sum_{h=1}^M g(\mathbf{p}_{T+1} | \mathbf{p}_{(h)}^T, \tau_{(h)}^*),$$

où  $\mathbf{p}_{T+1} | \mathbf{p}_T, \tau_2 \sim \text{Dirichlet}(\mathbf{p}_T, \tau_2)$ . Donc, d'après (22), (23) et (25), l'estimation et la prédiction sont simples. Par exemple, considérons le modèle

## 4. Analyse des données

variable en fonction du temps. Dans le cas de l'estimation, pour chaque  $h$ , nous tirons de (24) une valeur aléatoire  $\mathbf{p} | \mathbf{p}_{(h)}^{T-1}, \tau_{(h)}^* \sim \text{Dirichlet}(\mathbf{p}_{(h)}^{T-1}, \tau_{(h)}^*)$ , dénotée par  $\mathbf{p}_{(h)}^T$ ,  $h = 1, \dots, M$ . Pour la prédiction, pour chaque  $h$ , nous tirons de (25) une valeur aléatoire  $\mathbf{p}_{T+1} | \mathbf{p}_{(h)}^T, \tau_{(h)}^* \sim \text{Dirichlet}(\mathbf{p}_{(h)}^T, \tau_{(h)}^*)$ , dénotée par  $\mathbf{p}_{(h)}^{T+1}$ ,  $h = 1, \dots, M$ . Donc, l'inférence au sujet de  $\mathbf{p}_T$  et  $\mathbf{p}_{T+1}$  est conduite de la manière habituelle. La méthode est semblable pour les modèles de non-réponse ignorable et non ignorable.

Dans cette section, nous comparons nos modèles à ceux de Chen et Siasny (2003) et aux résultats réels (élection de novembre). Nous avons introduit un nouveau paramètre pour faciliter la prédiction du résultat de l'élection. Nous étudions aussi en détail la sensibilité de l'inférence au choix de  $\kappa_1$  et de  $\kappa_2$ . En nous fondant sur notre méthode, nous avons spécifié le coefficient de variation,  $c_0 = 0,031$ , et la moyenne,  $\mu_0 = 2,431$ , de la distribution *a priori* de  $\tau_1$ . Au tableau 2, nous comparons l'inférence au sujet des proportions d'électeurs d'octobre affectées aux trois candidats d'après nos modèles et d'après ceux de Chen et Siasny (2003). Dans ce tableau, les résultats sont fondés sur le prior  $\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2)$  (c'est-à-dire  $\kappa_1 = 1$ ). Nous présentons aussi les proportions réelles tirées de Chang et Krosnick (2001). Les proportions réelles sont (0,45, 0,50, 0,05) pour Fisher, Taft et les autres. En utilisant notre modèle de non-réponse variable en fonction du temps, nous estimons que ces proportions sont (0,41, 0,50, 0,09). Ces valeurs se comparent bien aux résultats réels. Les estimations correspondantes sont (0,41, 0,51, 0,08) pour le modèle de non-réponse ignorable et (0,40, 0,50, 0,09) pour le résultat de Chen et Siasny (2003) est celui qu'ils obtiennent à l'aide de leur modèle D qui produit les estimations (0,42, 0,51, 0,07). Nous fournissons des intervalles de crédibilité à 95 % pour nos estimations, mais dans le contexte de l'approche de Chen et Siasny (2003), il est un peu plus difficile de fournir des intervalles comparables. En outre, au tableau 2, nous présentons les estimations des proportions prédites pour l'élection de novembre. Les prédicteurs ponctuels sont comparables aux estimations ponctuelles, sauf pour la proportion prédite pour Taft sous le modèle de non-réponse ignorable. Toutefois, comme prévu, les intervalles de crédibilité à 95 % pour les proportions prédites sont beaucoup plus larges. Par exemple, sous le modèle variable en fonction du temps, l'intervalle de crédibilité à 95 % pour la proportion d'électeurs votant pour Taft au moment du sondage d'octobre est (0,41, 0,60) et pour la prédiction, il est (0,21, 0,78). Par conséquent, bien que les estimations et les prédictions ponctuelles indiquent

[0, 1]. Il est alors nécessaire d'utiliser une approximation par grille dans un intervalle proche de la borne, ce que l'on peut obtenir par rationnement en examinant la sortie de l'échantillonneur à mesure que son exécution progresse. S'il est probable qu'un paramètre compris dans  $[0, 1]$  s'écarte de 0 ou de 1, alors la méthode par grille convient bien; cela est le cas pour les  $\mu_i$ . Cependant, lorsqu'un paramètre tel que  $\tau_1$  (qui peut être très grand) est transformé en  $\phi$  dans l'intervalle  $[0, 1]$ ,  $\phi$  peut être très grand (proche de 1). Si la valeur transformée est de l'ordre de 0,999, il est nécessaire d'ajuster la recherche par grille afin qu'elle se fasse dans un intervalle contenant 0,999. Cet ajustement doit se faire par rationnement; il faut examiner la sortie de  $\phi$  à mesure que progresse l'exécution de l'échantillonneur et ajuster l'intervalle en conséquence. Par exemple, si l'intervalle  $[0, 1]$  contient 100 points de grille équidistants tels que 0,01, 0,02, 0,03, ..., 0,99, et que la valeur du paramètre se situe vraisemblablement autour de 0,999, même si nous procédons à un tirage uniforme dans l'intervalle de grille choisi, ces points de grille ne seront pas très efficaces.

Nous avons exécuté l'échantillonneur de Metropolis-Hastings en tirant une valeur aléatoire de chacune des densités (15), (16), (17), (18), (19), (20) et (21) en tirant la procédure complète jusqu'à la convergence. Il s'agit d'un exemple de l'échantillonneur Metropolis-Hastings (Ritter et Tanner 1992). Nous avons tiré un échantillon des densités  $a_{posterior}$  correspondant aux modèles de non-réponse ignorable et non ignorable de la même manière. Pour tous les modèles, nous avons utilisé un échantillon de  $M = 1\,000$  tiré des densités  $a_{posterior}$  pour l'estimation et la prédiction. Nous avons surveillé la convergence de chaque l'algorithme en examinant les traces graphiques de chaque paramètre en fonction de l'ordre des itérations et étudié le coefficient d'autocorrélation. Nous avons utilisé un échantillonneur de Gibbs pour ajuster les modèles de non-réponse ignorable et non ignorable. Nous avons procédé à 1 000 itérations « d'apprentissage », puis nous avons sélectionné une itération sur dix. Cette procédure donne de bons résultats.

Cependant, pour le modèle variable en fonction du temps, nous avons utilisé une approximation par grille de la loi  $a_{posterior}$  conditionnelle de  $p_0$ , mais les pas de Metropolis pour  $p_i, i = 1, \dots, T$ . Les pas de Metropolis n'ont pas donné de bons résultats, parce que les probabilités de saut étaient 0,67, 0,65 et 0,73 pour les trois densités  $a_{posterior}$  conditionnelles de  $p_1, p_2$  et  $p_3$ , alors qu'il est recommandé qu'elles soient comprises entre 0,25 et 0,50 (Gelman, Roberts et Gilks 1996); les réglages fins n'ont produit aucune amélioration. Par conséquent, nous avons également utilisé des approximations par grille de ces trois densités  $a_{posterior}$  conditionnelles. Ces approximations sont très précises. Dans chaque cas, nous avons commencé

### 3.4 Spécification de $\mu_0$ et $c_0$

Enfin, nous décrivons comment spécifier  $\mu_0$  et  $c_0$  dans (12). Cette étape est importante, parce qu'elle nous permet de « centrer » le modèle de non-réponse non ignorable sur le modèle de non-réponse ignorable (c'est-à-dire, un modèle d'expansion). Cette procédure s'inscrit dans la foulée de Nandram et coll. (2005).

Nous avons tiré un échantillon de  $\pi_{(h)}^i, i = 1, \dots, T, h = 1, \dots, M, M = 1\,000$  itérations à partir du modèle de non-réponse ignorable et calculé  $\pi_{(h)}^i = \sum_{j=1}^T \pi_{(h)}^{ij} / T, h = 1, \dots, M$ . Puis, en utilisant l'échantillonneur de Gibbs, nous avons ajusté le modèle

$$\pi_{(h)}^{iii} \sim \text{Dirichlet}(\mu_1 \tau_1),$$

$$\mu_1 \sim \text{Dirichlet}(1), p(\tau_1) = 1/(1 + \tau_1)^2, \tau_1 > 0,$$

où  $\mu_1$  et  $\tau_1$  *a priori* sont indépendants, pour obtenir un échantillon  $\tau_{(h)}^i, h = 1, \dots, M$ . Nous avons tiré 1 500 itérations avec un « apprentissage » de 500 pour obtenir  $M = 1\,000$  itérations.

Enfin, en prenant  $a = M^{-1} \sum_{h=1}^M \tau_{(h)}^i$  et  $b = (M-1)^{-1} \sum_{h=1}^M (\tau_{(h)}^i - a)^2$ , nous avons fixé que

$$c_0 = \sqrt{b/a} \text{ et } \mu_0 = a.$$

Pour les données électorales, notre méthode donne  $c_0 = 0,031$  et  $\mu_0 = 2,431$ . Cette spécification maintiendra le modèle de non-réponse non ignorable proche du modèle de non-réponse ignorable, fournissant par conséquent un mécanisme possible de centrage.



particulier, en posant que  $\mathbf{p} = (p_{ijk}^s, i = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$  et  $\boldsymbol{\pi} = (\pi_{ijk}^s, i = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ ,

$$y_{12j} | \{y_{1i}, p, \boldsymbol{\pi}\} \underset{\text{ind}}{\sim} \text{Multinomial}(\eta_j^p, \mathbf{q}^{(2)}), \quad j = 1, \dots, r,$$

$$y_{3k} | \{y_{1i}, p, \boldsymbol{\pi}\} \underset{\text{ind}}{\sim} \text{Multinomial}(v_k, \mathbf{q}^{(3)}), \quad k = 1, \dots, c,$$

$$y_{14} | \{y_{1i}, p, \boldsymbol{\pi}\} \sim \text{Multinomial}(w_i, \mathbf{q}^{(4)}), \quad (15)$$

où  $q_{ijk}^{(2)} = \pi_{12jk} / \sum_{k=1}^c \pi_{12jk}$ ,  $p_{ijk}^{(3)} = \pi_{13jk} / \sum_{j=1}^r \pi_{13jk}$  et  $q_{ijk}^{(4)} = \pi_{14jk} / \sum_{j=1}^r \sum_{k=1}^c \pi_{14jk}$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ ,  $i = 1, \dots, T$ . La densité *a posteriori* conditionnelle de  $\boldsymbol{\pi}^{ijk}$  est donnée par

$$\boldsymbol{\pi}^{ijk} | \{\mathbf{y}_1, \mathbf{y}_2, y_{1i}\} \underset{\text{ind}}{\sim} \text{Dirichlet}(y_{1i} + \mu_{11} \tau_{1i}, y_{12jk} + \mu_{12} \tau_{1i}, y_{13jk} + \mu_{13} \tau_{1i}, y_{14jk} + \mu_{14} \tau_{1i}) \quad (16)$$

avec indépendance sur  $i = 1, \dots, T$ ,  $j = 1, \dots, r$ ,  $k = 1, \dots, c$ . La densité *a posteriori* conditionnelle pour  $p_i$ ,  $i = 1, \dots, T$  est plus compliquée. Nous notons que

$$\pi(\mathbf{p}_0 | \sinon, y_1) \propto \prod_{j=1}^r \prod_{k=1}^c p_{ijk}^{p_{0jk} - 1} D(p_0 \tau_2) \quad (17)$$

et

$$\pi(p_i | \sinon, y_1)$$

$$\propto \left\{ \prod_{j=1}^r \prod_{k=1}^c p_{ijk}^{y_{1j,k} + p_{i,j,k} \tau_{2-1}} D(p_i \tau_2) \right\}, \quad i = 1, \dots, T, \quad (18)$$

où « *sinon* » fait référence à tous les paramètres compris dans  $(p, \boldsymbol{\pi}, \mu_1, \tau_1, \tau_2, y_1^{(1)})$  sauf  $p_0$  dans  $(17)$  et de (18) à l'annexe A.

Ensuite, nous considérons les hyperparamètres. En posant que  $\delta_s = \prod_{j=1}^r \prod_{k=1}^c \pi_{sjk}^{\delta_{sjk}}$ , et  $\boldsymbol{\pi} = (\pi_{ijk}^s, i = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ , la densité *a posteriori* conditionnelle conjointe de  $\mu_1, \tau_1$  est

$$p(\mu_1, \tau_1 | \boldsymbol{\pi}) \propto \frac{D(\mu_1 \tau_1)^{\{rcT\}}}{\prod_{\mu_1, \tau_1} \delta_{\mu_1, \tau_1}^{s-1}} \tau_1^{1/c\delta-1} e^{-\tau_1 / \mu_1 \delta \delta_s},$$

où

$$\sum_{s=1}^4 \mu_{1s} = 1, \mu_{1s} \geq 0, s = 1, 2, 3, 4, \tau_1 > 0.$$

Il n'est pas nécessaire d'obtenir un échantillon directement de  $p(\mu_1 | \tau_1, \boldsymbol{\pi})$ . Toutefois, en représentant par  $\boldsymbol{\mu}^{(s)}$  le vecteur de toutes les composantes de  $\mu_1$  sauf  $\mu_{1s}$ , nous avons

$$0 \leq \mu_{1s} \leq 1 - \sum_{s=1, s \neq 3}^4 \mu_{1s}, \quad s = 1, 2, 3. \quad (19)$$

Nous utilisons une méthode de grille pour tirer un échantillon à partir de  $p(\mu_{1s} | \mu_1^{(s)}, \tau_1, \boldsymbol{\pi})$ . Nous commençons par utiliser 50 mailles (autrement dit, nous divisons l'étendue de  $\mu_{1s}$   $(0, 1 - \sum_{s=1, s \neq 3}^4 \mu_{1s})$ , en 50 intervalles égaux) pour former une fonction de masse de probabilité approximative de  $\mu_{1s}$ ,  $s = 1, 2, 3$ . Nous tirons d'abord une variable aléatoire de la fonction de masses de probabilité pour indiquer lequel des 50 intervalles est sélectionné. Puis, pour  $\mu_{1s}$ , nous tirons une variable aléatoire uniforme dans cet intervalle. Cette procédure est efficace, parce que  $\mu_{1s}$  est bornée, que les intervalles sont très étroits et qu'il est très « peu coûteux » de construire la fonction de masse de probabilité discrète pour chaque  $\mu_{1s}$ ,  $s = 1, 2, 3$ . Enfin, nous obtenons  $\mu_{14}$  à partir de sa densité *a posteriori* conditionnelle en prenant  $\mu_{14} = 1 - \sum_{s=1}^3 \mu_{1s}$ .

La densité *a posteriori* conditionnelle de  $\tau_1$  est

$$p(\tau_1 | \mu_1, \boldsymbol{\pi}) \propto \left[ \prod_{s=1}^4 \frac{\Gamma(\mu_{1s} \tau_1)^{\{rcT\}}}{\delta_{\mu_{1s} \tau_1}^{s-1}} \right] \tau_1^{1/c\delta-1} e^{-\tau_1 / \mu_1 \delta \delta_s}, \quad \tau_1 > 0. \quad (20)$$

Pour tirer une valeur aléatoire de (20), nous procédons de la même manière que pour (19), sauf que nous transformons  $\tau_1$  de la moitié positive de la droite réelle en  $(0, 1)$ . (Il est plus commode d'obtenir une approximation par grille d'une densité dans un intervalle borné.) Donc, en posant que  $\tau_1 = \phi/(1 - \phi)$  dans (20), nous avons

$$p(\phi | \mu_1, \boldsymbol{\pi})$$

$\phi/(1 - \phi)$ . En posant que  $p = (p_{ijk}^s, i = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c)$ , la densité *a posteriori* conditionnelle de  $\tau_2$  est

$$\pi(\tau_2 | p) \propto \frac{(1 + \tau_2)^2}{\prod_{j=1}^r \prod_{k=1}^c D(p_{jk}^{1-\tau_2})} \left\{ \frac{D(p_{jk}^{1-\tau_2})}{\prod_{j=1}^r \prod_{k=1}^c D(p_{jk}^{1-\tau_2})} \right\}, \quad \tau_2 > 0. \quad (21)$$

Nous obtenons un échantillon de manière comparable à  $\tau_1$  dans (20). Nous possédons une vaste expérience de l'utilisation de l'approximation par grille. Cependant, il faut être prudent lorsque l'on applique cette méthode pour des paramètres dont les valeurs sont proches de 0 ou de 1 dans l'intervalle

$$E\{p_i | p_{-i-1}, t = 1, \dots, T + 1;$$

de sorte que  $\{p_i\}$ , *a priori*, est une martingale vectorielle. Ici,  $T$  est petit (c'est-à-dire  $T = 3$ ). Donc, cette structure

variable en fonction du temps semble plus appropriée et pourrait accroître la précision. Notons également que nous

avons pris  $p_0 \sim \text{Dirichlet}(1)$ .

Enfin, nous spécifions les densités *a priori* pour les hyperparamètres. Premièrement, nous prenons

$$\mu_1, \mu_2 \sim \text{Dirichlet}(1), \quad (10)$$

c'est-à-dire des densités *a priori* essentiellement

informatives.

Enfin,  $\tau_1$  et  $\tau_2$  sont des variables aléatoires indé-

pendantes et identiquement distribuées provenant de

$$f(x) = 1/(1+x)^2, \quad x \geq 0. \quad (11)$$

De nouveau, il s'agit d'une densité *a priori* essentiellement

non informative. Notons que  $\mu_1$  et  $\tau_1$  n'existent pas dans le

modèle de non-réponse ignorable. Gelman (2006) a

recommandé d'utiliser des lois *a priori* tels que (11) au lieu

des priors gamma diffus appropriés dont le comportement

est mauvais.

Pour les modèles de non-réponse non ignorable, nous

devons spécifier plus prudemment la densité *a priori* de  $\tau_1$ ,

parce que les  $\pi_{ijk}$  ne sont pas identifiables. Ici, nous

essayons de «centrer» les modèles de non-réponse non

ignorable sur le modèle de non-réponse ignorable. Dans (8),

le paramètre  $\tau_1$  nous renseigne sur la mesure dans laquelle

le modèle de non-réponse non ignorable est proche du

petit, les  $\pi_{ijk}$  seront très différentes et si  $\tau_1$  est grand, les

$\pi_{ijk}$  seront fort semblables. Donc, l'inférence *a priori* sera

sensible au choix de  $\tau_1$ , qui doit donc être fait avec

prudence. Nous aimerions choisir pour  $\tau_1$  une densité

*a priori* telle que le modèle de non-réponse non ignorable

reste proche du modèle de non-réponse ignorable. Donc,

nous prenons

$$\tau_1 \sim \text{Gamma}(1/c_0^2, 1/\mu_0 c_0^2), \quad (12)$$

où  $E(\tau_1) = \mu_0$  et  $\text{CV}(\tau_1) = c_0$ , CV étant le coefficient de

variation; nous devons spécifier  $\mu_0$  ainsi que  $c_0$ . Nous

utilisons la loi *a priori* (12), parce que, par un choix

approprié de  $\mu_0$  et  $c_0$ , il est possible de centrer le modèle

de non-réponse non ignorable sur le modèle de non-réponse

ignorable. Naturellement, d'autres lois *a priori* appropriées

comme des comprenant des paramètres tels que  $\mu_0$  et  $c_0$

peuvent être utilisées pour faciliter le centrage. À la section

3.4, nous utiliserons des échantillons provenant de la densité

*a posteriori* de  $\tau_1$  sous le modèle de non-réponse ignorable

pour spécifier  $\mu_0$  et  $c_0$ .

Pour chacun des trois modèles, il est facile d'écrire la densité *a priori* conjointe des paramètres. Par exemple, pour le modèle variable en fonction du temps, la densité *a priori* conjointe est

$$p(p, \pi, \mu_1, \tau_1, \tau_2) \propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{(1 + \tau_2)^2}{1} \left\{ \prod_{i=1}^T \prod_{j=1}^c \prod_{k=1}^{p_{i-1}/k - \tau_2 - 1} D(p_{i-1}, \tau_2) \frac{\prod_{s=1}^j \prod_{c=1}^k \prod_{i=1}^4 \pi_{tsjk}^{u_{tsjk} - \tau_2 - 1}}{D(\mu_1, \tau_1)} \right\}, \quad (13)$$

où  $D(\cdot)$  est la fonction de Dirichlet.

### 3.3 Ajustement du modèle de non-réponse non ignorable variable en fonction du temps

Si nous combinons la fonction de vraisemblance (5) et la densité *a priori* conjointe (13) par la voie du théorème de Bayes, nous obtenons la densité *a posteriori* conjointe des paramètres  $\pi$ ,  $p$ ,  $\mu_1$ ,  $\tau_1$ ,  $\tau_2$  et des variables latentes  $y^{(i)}$  qui

suit

$$\pi(p, \pi, \mu_1, \tau_1, \tau_2, y^{(i)} | y^{(i)})$$

$$\propto \tau_1^{1/c_0^2 - 1} e^{-\tau_1/\mu_0 c_0^2} \frac{1}{(1 + \tau_2)^2} \prod_{i=1}^T \prod_{j=1}^c \prod_{k=1}^{p_{i-1}/k - \tau_2 - 1} \left[ \prod_{s=1}^j \prod_{c=1}^k \prod_{i=1}^4 \pi_{tsjk}^{u_{tsjk}} \right] \left[ \prod_{s=1}^j \prod_{c=1}^k \prod_{i=1}^4 \pi_{tsjk}^{v_{tsjk}} \right] \left\{ \prod_{i=1}^T \prod_{j=1}^c \prod_{k=1}^{p_{i-1}/k - \tau_2 - 1} \frac{D(p_{i-1}, \tau_2)}{\prod_{s=1}^j \prod_{c=1}^k \prod_{i=1}^4 \pi_{tsjk}^{u_{tsjk} - \tau_2 - 1}} D(\mu_1, \tau_1) \right\} \quad (14)$$

sachant  $\sum_{k=1}^c y_{i2jk} = u_{ij}, j = 1, \dots, T, \sum_{j=1}^c y_{i3jk} = v_{ik}, k = 1, \dots, c$ , et  $\sum_{j=1}^c \sum_{k=1}^c y_{i4jk} = w_i, i = 1, \dots, T$ .

La densité *a posteriori* (14) est complexe, de sorte que nous utiliserons des méthodes de Monte Carlo par chaîne de

Markov pour l'ajuster. Cependant, il est facile d'ajuster le

modèle variable en fonction du temps en utilisant l'échan-

tilleonneur matriciel de Metropolis-Hastings comme nous le

décrivons. En outre, de manière semblable, en utilisant

l'échantillonneur de Gibbs (Ritter et Tanner 1992), il est

facile d'ajuster les modèles de non-réponse ignorable et non

ignorable. Nous tirons un échantillon à partir de la densité

*a posteriori* conjointe afin de conduire une inférence au

sujet des paramètres. Plus précisément, nous devons faire

une inférence au sujet de  $p$ . Pour exécuter l'échan-

tilleonneur de Metropolis-Hastings, nous avons besoin de la

densité *a posteriori* conditionnelle de chacun des para-

mètres, sachant les autres.

En premier lieu, nous considérons les fonctions de masse

de probabilité *a posteriori* conditionnelles de  $y_{is}, s =$

2, 3, 4,  $i = 1, \dots, T$ , sachant  $y^{(s)}, j = 1, \dots, r, k =$

1,  $\dots, c$ . De (14), il découle clairement que, sous la densité

*a posteriori* conditionnelle, les  $y_{is}, i = 1, \dots, T, s = 2, 3, 4$ ,

sont des vecteurs aléatoires multinomiaux indépendants. En

### 3.2 Modèles de non-réponse

En posant que  $\mathbf{I}^{(t)} = (I^{(t)k}, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c, \ell = 1, \dots, L)$ , pour tous les modèles, nous prenons

$$(1) \quad \mathbf{I}^{(t)} | \mathbf{p}_t \sim \text{Multinomial}(\{\mathbf{p}_t\},$$

où

$$\sum_{k=1}^c \sum_{j=1}^r p_{ijk} = 1, p_{ijk} \geq 0, t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c,$$

Pour le modèle de non-réponse ignorable, nous prenons

$$(2) \quad \mathbf{I}^{(t)} | \pi_t \sim \text{Multinomial}(\{\pi_t\},$$

Autrement dit, il n'y a aucune dépendance à l'égard de la situation de cellule d'un individu. Pour les modèles de non-réponse non ignorable, nous prenons

$$\mathbf{I}^{(t)} | \{I^{(t)k\ell} = 1, I^{(t)k\ell} = 0, j \neq j', k \neq k', \pi_{jk}^{(t)}\}$$

$$(3) \quad \sim \text{Multinomial}(\{\pi_{jk}^{(t)}\},$$

L'hypothèse (3) spécifie que les probabilités qu'un individu appartienne à l'un des quatre tableaux dépend des deux caractéristiques (c'est-à-dire classifications de ligne et de colonne) de cet individu. De cette manière, nous

intégrons l'hypothèse que les données manquantes sont non ignorables. Notons, sachant les paramètres spécifiés dans (1) à (3), que le comportement d'un électeur est corréli à celui d'un autre à un même temps  $t$ , mais que ces comportements sont indépendants au cours du temps. Il mérite d'être souligné ici que les paramètres de (2) sont identifiables, mais que ceux de (3) ne le sont pas. C'est là que survient la difficulté dans le modèle de non-réponse non ignorable et il convient d'y accorder une attention particulière.

Il découle de (1) et de (2) que, pour le modèle ignorable,

$$(4) \quad \mathcal{G}(\mathbf{p}, \pi | \mathbf{y}) \propto \prod_{t=1}^T \prod_{j=1}^r \prod_{k=1}^c \pi_{jk}^{I_{tjk}} \left[ \prod_{s=1}^s \prod_{j=1}^r \prod_{k=1}^c \prod_{\ell=1}^L p_{jks\ell}^{I_{tjks\ell}} \right]$$

sachant  $\sum_{k=1}^c \sum_{j=1}^r Y_{12jk} = u_j, j = 1, \dots, r, \sum_{j=1}^r \sum_{k=1}^c Y_{13jk} = v_k, k = 1, \dots, c$  et  $\sum_{j=1}^r \sum_{k=1}^c \sum_{\ell=1}^L Y_{14jk\ell} = w_\ell$ . Notons que, sous des conditions d'ignorabilité, la fonction de vraisemblance dans (4) se divise en deux parties, l'une contenant les  $\pi_{jk}$  uniquement et l'autre, les  $p_{jks\ell}$  et que les inférences au sujet de ces paramètres ne sont pas liées; voir la section 3.2 de Nandram et coll. (2005) pour la discussion originale de ce modèle. En outre, il découle de (1) et de (3) que, pour les modèles de non-réponse non ignorable, la fonction de

vraisemblance augmentée pour  $\mathbf{p}, \pi, \mathbf{y} | \mathbf{y}^{(0)}$  est

$$(5) \quad \mathcal{G}(\mathbf{p}, \pi, \mathbf{y} | \mathbf{y}^{(0)}) \propto \prod_{t=1}^T \prod_{j=1}^r \prod_{k=1}^c \pi_{jk}^{I_{tjk}} \left[ \prod_{s,j,k} p_{jks\ell}^{I_{tjks\ell}} \right] \prod_{r,c} p_{ijk}^{I_{tijk}}$$

où  $\mathbf{p}_0$  est également connue. Notons que

$$(9) \quad \mathbf{p}_t | \mathbf{p}_{t-1}, \tau_2 \sim \text{Dirichlet}(\mathbf{p}_{t-1}, \tau_2), t = 1, \dots, T + 1,$$

temps, nous prenons

Pour le modèle de non-réponse variable en fonction du ignorable sur le modèle de non-réponse ignorable.

différemment est de « centrer » le modèle de non-réponse non d'information » au cours du temps. Fait plus important encore, parce que les  $\pi_{jk}$  ne sont pas identifiables, il en est D'abord, nous notons que (8) fournit un « emprunt

$$(8) \quad \sim \text{Dirichlet}(\mu_1 \tau_1), t = 1, \dots, T, j = 1, \dots, r, k = 1, \dots, c,$$

$$\pi_{jk} | \mu_1, \tau_1$$

nous prenons

données. Pour les modèles de non-réponse non ignorable, loi de Dirichlet en ne disposant pratiquement d'aucune sinon, nous devrions spécifier les paramètres inconnus de la  $a priori$  uniforme dans  $R^4$  (essentiellement non informatif);  $T = 3$  dans notre application). Donc, nous utilisons la loi Nous avons besoin de (7), parce que  $T$  est petit (c'est-à-dire où  $\mathbf{I}$  est un vecteur de dimension quatre de valeurs unitaires.

$$(7) \quad \pi_t \sim \text{Dirichlet}(\mathbf{1}), t = 1, \dots, T,$$

prenons

$\pi_{jk}$ . Pour le modèle de non-réponse ignorable, nous Décritons maintenant les modèles stochastiques pour les

nomiale dans nos travaux de recherche. avons utilisé extensivement la combinaison Dirichlet-Multi-notre premier article (c'est-à-dire Nandram 1998), nous dans l'échantillonnage multinomial; depuis la publication de exactement, son utilisation en tant que loi  $a priori$  conjuguée ce qui concerne la distribution dans l'urne de Polya et, plus consulter Ghosh et Mleden (1997, pages 42, 50 et 127) en Dirichlet et  $\alpha_j > 0, j = 1, \dots, k$ . Pour une référence rapide,  $x_j \geq 0, j = 1, \dots, k, \sum_{j=1}^k x_j = 1$ , où  $D(\alpha)$  est la fonction de suit une loi de Dirichlet si  $p(\mathbf{x} | \alpha) = \prod_{j=1}^k x_j^{\alpha_j - 1} / D(\alpha)$ , cours du temps. Notons que le vecteur  $\mathbf{x}$  de dimension  $k$  probabiliste de (6) permet un « emprunt d'information » au c'est-à-dire une étape plus loin (novembre). La structure où nous considérons la prédiction au temps  $T + 1$ ,

$$(6) \quad \mathbf{p}_t | \mu_2, \tau_2 \sim \text{Dirichlet}(\mu_2 \tau_2), t = 1, \dots, T + 1,$$

ignorable, nous prenons

Pour les modèles de non-réponse ignorable et non semblable.

pour une description de l'identifiabilité dans une situation 1, ..., c, et  $\sum_{k=1}^c \sum_{j=1}^r Y_{14jk} = w_\ell$ ; voir Nandram et coll. (2005) sachant  $\sum_{k=1}^c \sum_{j=1}^r Y_{12jk} = u_j, j = 1, \dots, r, \sum_{j=1}^r \sum_{k=1}^c Y_{13jk} = v_k, k =$





Nous décrivons brièvement le tableau de contingence  $2 \times 3$  des données électorales en Ohio selon l'intention de voter (IV) et le candidat préféré (CAN). Ici, IV est une variable binaire comportant deux niveaux, à savoir votera vraisemblablement et ne votera vraisemblablement pas; CAN comporte trois niveaux, à savoir Fisher, Taft et autre. Les variables IV et CAN comprennent aussi des électeurs indécis. La majeure partie de ceux-ci proviennent des groupes d'électeurs qui « voteront vraisemblablement » et « ne voteront vraisemblablement pas », et les chiffres sont de 173, 142 et 138 pour janvier, avril et octobre, respectivement; les nombres d'électeurs indécis pour Fisher, Taft et autre sont beaucoup plus faibles.

Dans le sondage de janvier 1998, environ 73 % des électeurs sont complètement classés, 27 % n'ont pas décidé de leur candidat préféré, 1 % seulement ne savent pas s'ils voteront ou non et cinq personnes seulement sont entièrement non classées parmi les 648 participants. L'ensemble de données utilisé dans notre étude est présenté au tableau 1 sous forme de tableau de contingence  $2 \times 3$  de l'intention de voter et du candidat préféré. Notre problème est de prédire le candidat gagnant en estimant la proportion de votes finaux pour chaque candidat.

Tableau 1  
Classification des données du sondage de l'État d'Ohio  
d'octobre 1998 selon l'intention de voter et le candidat

Candidate	Fisher	Taft	Autre	Indécis	Total	Intention
-----------	--------	------	-------	---------	-------	-----------

a. Janvier 1998		b. Avril 1998		c. Octobre 1998	
Votera	127	183	8	109	427
vraisemblablement	127	183	8	109	427
Ne votera	57	94	4	59	214
vraisemblablement	57	94	4	59	214
pas	0	2	0	5	7
Indécis	184	279	12	173	648
Total	114	135	1	61	311
Votera	114	135	1	61	311
vraisemblablement	114	135	1	61	311
Ne votera	104	149	3	78	334
vraisemblablement	104	149	3	78	334
pas	2	6	0	3	11
Indécis	220	290	4	142	656
Total	112	140	23	61	336
Votera	112	140	23	61	336
vraisemblablement	112	140	23	61	336
Ne votera	96	108	21	73	298
vraisemblablement	96	108	21	73	298
pas	7	11	1	4	23
Indécis	215	259	45	138	657
Total	215	259	45	138	657

NOTA : Ces données sont tirées de Chang et Krosnick (2001); Chen et Stasny (2003) ont utilisé un ensemble de données fort semblable; ils n'ont pas utilisé 5, 3 et 4, c'est-à-dire le nombre d'électeurs indécis dans les deux variables.

Nandram et Choi : Une répartition bayésienne des électeurs indécis

Les échantillons obtenus en janvier, en avril et en octobre sont indépendants. Il n'y a pas de suréchantillonnage pour une sous-population particulière ni de pondération de « télephoniques, la base de sondage par composition aléatoire (CA) souffre souvent du problème de sous-décombrement. Comme la couverture téléphonique n'est pas uniforme selon l'âge, la race, le sexe, le revenu et la géographie, il est nécessaire de poststratifier l'échantillon original afin de réduire le biais de couverture par pondération correcte des données originales.

Nous exécutons un test préliminaire de l'hétérogénéité des proportions de cellule entre les trois sondages. En supposant que les données manquent au hasard, nous ajoutons les électeurs indécis. Nous supposons que, pour chaque ligne (colonne), les électeurs indécis sont ajoutés proportionnellement aux effectifs de cellule. Soit  $n_{ijk}^*$  les effectifs corrigés de cellule avec  $n_i^* = \sum_{j=1}^J \sum_{k=1}^K n_{ijk}^*$  et soit  $d_{ijk}^*$  les proportions de cellule. Pour un modèle à proportions hétérogènes, nous nous attendons

où  $\mathbf{I}$  est un vecteur  $nc$  de valeurs unitaires.

Alors, le facteur de Bayes d'hétérogénéité est

$$FB = \frac{1}{\prod_{j=1}^J \prod_{k=1}^K \prod_{c=1}^C \left[ (\rho c - 1)^{\eta_2} \right]} \left[ \sum_{t=1}^{T-1} u^{g/k}_t i \right] \left[ \sum_{t'=1}^{T-1} u^{g/k}_{t'} i \right] \left[ \sum_{t''=1}^{T-1} u^{r/c}_{t''} i \right]$$

### 3. Méthodologie

Nous avons construit un modèle de non-réponse non ignorable variable en fonction du temps pour les données du sondage de l'Etat de l'Ohio de 1998. Aux fins de comparaison, nous avons également envisagé deux autres

la discussion dans Kalton et Kasprzyk (1986). Pourtant, la non-réponse totale pourrait également être non ignorable et il serait alors nécessaire d'inclure simultanément les non-réponses partielle et totale.

Dans le présent article, notre principale contribution consiste à introduire une méthode bayésienne pour analyser des données provenant d'un tableau de contingence  $r \times c$ , en présence à la fois de non-réponse partielle et totale, et dans des conditions où le mécanisme de création des données manquantes peut être non ignorable avec une structure variable en fonction du temps. À la section 2, nous décrivons les données catégoriques sur l'intention de voter et le candidat préféré avec une structure variable en fonction du temps. À la section 3, nous exposons la méthode appliquée pour estimer les probabilités de cellule en intégrant les deux types de données manquantes, et nous montrons comment étendre un modèle de non-réponse ignorable à un modèle de non-réponse non ignorable et à un modèle de non-réponse variant en fonction du temps. Nous montrons aussi comment utiliser les méthodes de Monte Carlo par chaîne de Markov pour ajuster le modèle de non-réponse non ignorable. À la section 4, nous analysons les données des sondages électoraux réalisés en Ohio pour démontrer la souplesse de nos méthodes. Enfin, à la section 5, nous présentons nos conclusions.

## 2. Données des sondages électoraux de 1998 en Ohio

Le Center for Survey Research (CSR) de la Ohio State University a réalisé le Buckeye State Poll (BSP), ou sondage électoral dans l'État de l'Ohio, durant les élections de 1998 en vue de désigner le sénateur, le gouverneur, le procureur général, le secrétaire d'État, le trésorier et le maire de Columbus. Au cours de certains mois avant les élections, le CSR a réalisé des sondages préélectoraux dans le cadre du BSP et a inclus des questions supplémentaires pour recueillir des renseignements sur la probabilité de voter et sur le candidat préféré. Dans le BSP, les ménages sont échantillonnés par la méthode de composition aléatoire (CA) et un adulte par ménage est sélectionné pour l'interview par la méthode du dernier anniversaire (Lavarakas 1993).

Suit une brève description de la méthode CA. Les entreprises de sondage y recourent à grande échelle, avant tout en vue d'obtenir un échantillon représentatif de la population globale d'électeurs. L'échantillonnage par la méthode CA repose sur l'hypothèse qu'il n'est pas possible d'obtenir un échantillon représentatif en utilisant les numéros de téléphone répertoriés dans l'annuaire. Chaque numéro de téléphone contient dix chiffres, dont les trois

premiers forment l'indicateur régional, les trois suivants forment le préfixe (aussi appelé indicatif de central), tandis que les quatre derniers (suffixe) identifient un abonné particulier ou un ménage (un ménage peut avoir plus d'un numéro de téléphone). Les codes régionaux sont de nature géographique et identifient habituellement les localités dans un État, et les indicatifs de central peuvent aussi avoir une orientation géographique. Il existe dix millions de numéros à composer, mais approximativement moins de 25 % de ceux-ci sont des numéros de téléphone réels. Par conséquent, du temps et de l'argent sont gaspillés en composant des numéros non attribués. Nous discutons de ce point plus en détail à la section 3.

Chen et Stasny (2003), ainsi que Chang et Krosnick (2001) ont analysé les données provenant de trois sondages préélectoraux du BSP. Des détails sur chacun de ces trois sondages figurent au tableau 1. Dans le cadre de ces sondages, on a déterminé trois fois le candidat préféré des répondants (en janvier, en avril et en octobre) pour la course au poste de gouverneur de l'Ohio de novembre 1998. En outre, on a posé deux questions aux répondants pour déterminer la probabilité qu'ils se rendent aux urnes le jour de l'élection. Chang et Krosnick (2001) ont également utilisé des variables filtres (telles que l'inscription sur la liste des électeurs, la probabilité autodéclarée de voter et le fait d'avoir voté lors de la dernière élection importante, etc.) pour déterminer quelles personnes étaient le plus susceptibles de voter. Donc, la prédiction est fondée uniquement sur les répondants susceptibles de voter. Les personnes inscrites sur les listes d'électeurs sont classées en trois catégories, à savoir votera probablement, ne votera probablement pas et indécis. Chang et Krosnick (2001) ont montré que la répartition déterministe des répondants indécis améliore la prévision des préférences des électeurs pour les candidats, comparativement à l'exclusion de tous les répondants indécis. Chen et Stasny (2003) ont utilisé des modèles probabilistes pour répartir les électeurs indécis et ont comparé leurs prévisions à celles de Chang et Krosnick (2001).

L'ensemble de données de Chen et Stasny (2003) diffère légèrement du nôtre, parce que nous utilisons les nombres d'indécis (non-réponse totale) sur les deux variables. Un électeur peut être indécis pour au moins une des deux variables catégoriques à chacun des trois sondages. Chen et Stasny (2003) n'étudient les données comprenant les indécis que pour exactement une variable, et non les deux. Dans le tableau 1, pour les électeurs indécis sur les deux variables, les chiffres pour les sondages de janvier, d'avril et d'octobre sont de 5, 3 et 4, respectivement; ils sont représentés en caractères gras. En fait, l'inclusion de ces dénominations dans notre modèle est une extension des modèles de Chen et Stasny qui généralise considérablement notre méthode.





# Une répartition bayésienne des électeurs indécis

Balgobin Nandram et Jai Won Choi<sup>1</sup>

## Résumé

Aux États-Unis, les données provenant des sondages électoraux sont habituellement présentées dans des tableaux de contingence à double entrée et de nombreux sondages sont réalisés avant qu'il n'y ait lieu l'élection réelle en novembre. Par exemple, pour l'élection du gouverneur de l'État de l'Ohio en 1998, trois sondages (Buckeye State Poll) ont eu lieu, un en janvier, un en avril et un en octobre; la première catégorie des tableaux représente les candidats (par exemple Fisher, Taft et autre) et la deuxième représente l'intention courante de vote (votera vraisemblablement ou ne votera vraisemblablement pas pour le gouverneur de l'Ohio). Le nombre d'électeurs indécis est important dans l'une ou dans les deux catégories pour les trois sondages et nous utilisons une méthode bayésienne pour les répartir entre les trois candidats. Nous pouvons ainsi modéliser divers scénarios de données manquantes sous les hypothèses d'ignorabilité et de non-ignorabilité, et nous utilisons un modèle Dirichlet-Multinomial pour estimer les probabilités de cellule qui nous aideront à prédire le gagnant. Nous proposons un modèle de non-réponse non ignorable variable en fonction du temps pour les trois tableaux. Ici, un modèle de non-réponse non ignorable est centré sur un modèle de non-réponse ignorable afin d'introduire une certaine souplesse et une certaine incertitude au sujet de l'ignorabilité ou de la non-ignorabilité. Nous considérons également deux autres modèles concurrents, à savoir un modèle de non-réponse ignorable et un modèle de non-réponse non ignorable. Ces deux derniers modèles reposent sur l'hypothèse d'un processus stochastique commun pour obtenir un renforcement par emprunt de données au cours du temps. Nous utilisons des méthodes de Monte Carlo par chaîne de Markov pour ajuster les modèles. Nous construisons aussi un paramètre qui peut éventuellement être utilisé pour prédire le gagnant parmi les candidats à l'élection de novembre.

Mots clés : Monte Carlo par chaîne de Markov; échantillonneur de Metropolis; modèle Dirichlet-Multinomial; modèle variable en fonction du temps; tableau de contingence à double entrée.

## 1. Introduction

Il est courant, en pratique, d'utiliser des tableaux de contingence à double entrée pour présenter les données de sondage. Notre application a pour but de prédire le gagnant d'une élection en utilisant des tableaux construits en se basant sur une courte série de sondages électoraux réalisés avant l'élection réelle. Dans le cas de nombreux sondages, des données manquent, si bien que la classification des individus échantillonnés n'est que partielle. Little et Rubin (2002, section 1.3) définissent les trois mécanismes de création de données manquantes [manquant complètement au hasard (MCAR pour « missing completely at random »), au hasard (MAR pour « missing at random »), ne manquent pas au hasard (MNAR pour « missing not at random »)]; des modèles à données manquantes ignorables sont utilisés pour analyser les données issues des mécanismes MAR et MCAR, et des modèles à données manquantes non ignorables sont utilisés pour les données provenant du mécanisme MNAR. Donc, pour le tableau à double entrée, il existe à la fois une non-réponse partielle (l'une des deux catégories manque) et une non-réponse totale (les deux catégories manquent). Comme le mécanisme de création des données manquantes pourrait ne pas être connu, il serait peut-être préférable d'utiliser un

modèle dans lequel est incluse une différence entre les données observées et les données manquantes (c'est-à-dire des données manquantes non ignorables). Pour un tableau de contingence général  $r \times c$ , nous examinons la question de l'estimation des probabilités de cellule du tableau à double entrée. Elle est importante, car, si le nombre d'électeurs indécis est considérable, la prédiction du résultat d'une élection fondée uniquement sur les données partiellement observées pourrait être incorrecte. Comme dans Nandram, Cox et Choi (2005), nous avons essentiellement quatre tableaux à double entrée, à savoir le tableau contenant les cas complets et trois tableaux supplémentaires. Le premier de ceux-ci ne contient que les classifications par ligne (non-réponse partielle), le deuxième ne contient que les classifications par colonne (non-réponse partielle) et le troisième ne contient aucune classification (non-réponse totale). Nous avons étendu les modèles de non-réponse ignorable et non ignorable établis pour les tableaux de contingence à double entrée par Nandram et coll. (2005) afin d'inclure une troisième catégorie (c'est-à-dire le temps dans une brève série de sondages électoraux). Nous avons étendu ces modèles encore davantage afin d'intégrer une structure de non-réponse non ignorable variable en fonction du temps, qui peut fournir une prédiction plus efficace. Une méthode bayésienne nous

En outre,

$$(A.11) \quad \sum_{i=1}^n \omega_i \hat{y}_i = \sum_{i=1}^n \omega_i Y_i$$

En combinant (A.8), (A.10) et (A.11), nous obtenons que l'égalité tient si, et uniquement si,  $\hat{y}_i = Y_i$ . Donc, nous avons montré que  $\hat{y}_i$  est le seul prédicteur linéaire sans biais qui satisfait (7) et minimise le critère (18).

## Remerciements

Cette étude a été financée en partie par l'entente de coopération 68-3A75-14 conclue entre le Natural Resources Conservation Service de l'USDA et la Iowa State University.

## Bibliographie

- Battese, G.E., Harter, R.M. et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of American Statistical Association*, 28-36.
- Harville, D.A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *Annals of Statistics*, 384-395.
- You, Y., et Rao, J.N.K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference*, 197-208.
- You, Y., et Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 431-439.
- Wang, J., et Fuller, W.A. (2003). The mean square error of small area estimators constructed with estimated area variances. *Journal of American Statistical Association*, 716-723.
- Robinson, G.K. (1991). That MPLSB is a good thing: the estimation of random effects. *Statistical Science*, 15-51.
- Robinson, G.K. (1991). *Small area estimation*. New York : John Wiley & Sons, Inc.
- Rao, J.N.K. (2003). *Small area estimation*. New York : John Wiley & Sons, Inc.
- Pfeiffermann, D., et Barnard, C.H. (1991). Some new estimators for small-area means with application to the assessment of Farmland values. *Journal of Business & Economic Statistics*, 73-84.
- Maniel, H.J., Singh, A.C., et Barreau, M. (1993). Benchmarking of small area estimators. Dans *Proceedings of International Conference on Establishment Surveys*, American Statistical Association, Washington, DC, 920-925.
- Isaki, C.T., Tsay, J.H. et Fuller, W.A. (2000). Estimation des facteurs de correction au recensement. *Techniques d'enquête*, 31-42.
- Henderson, C.R. (1963). Selection index and expected genetic advance. Dans *Statistical Genetics and Plant Breeding*, (Eds., W.D. Hanson et H.F. Robinson), National Academy of Sciences-National Research Council, Washington, Publication 982, 141-163.





La valeur du critère  $\hat{Q}$  pour MPLSBE est 1,20 sous le modèle (B). Donc, l'estimation des paramètres accroit la variance moyenne des prédicteurs d'environ 20 % comparativement à l'utilisation des paramètres connus. Si les prédictions sont faites en utilisant des valeurs connues de  $\sigma^2_{\epsilon}$ , la valeur du critère  $\hat{Q}$  pour MPLSBE est 1,06. Par conséquent, l'estimation de  $\sigma^2_{\epsilon}$  est celle qui contribue le plus à l'accroissement de la variabilité. Comme le biais est nul quand le modèle (B) est utilisé pour l'estimation, les ajustements que produisent les prédicteurs contraints sont faibles, si bien que les prédicteurs ajustés donnent des valeurs de critères fort semblables à celles des prédicteurs non ajustés. Les prédicteurs YR ont des valeurs de critère un peu plus grandes que les prédicteurs PB et ITF correspondants, parce que le prédicteur YR utilise un estimateur inefficace de  $\beta$ . Les prédicteurs reposant sur le modèle augmenté ont une valeur de critère  $\hat{Q}$  légèrement plus grande, parce que le modèle contient une variable redondante. L'accroissement de moins de 2 % de  $\hat{Q}$  est de l'ordre de  $n^{-1}$ , c'est-à-dire la perte attendue en raison de l'ajout d'un paramètre inutile dans une prédiction par les moindres carrés.

La valeur du critère  $\hat{Q}$  pour MPLSBE sous le modèle (A) est de 1,438 comparativement à 1,203 pour MPLSBE sous le modèle (B). Il s'agit de la pénalité due à l'erreur de spécification du modèle. Parmi les méthodes d'ajustement fondées sur le modèle (A), la méthode ITF est celle dont la valeur de  $\hat{Q}$  la plus faible. Les différences entre les prédicteurs PB, ITF et MPLSBE sont faibles. La valeur de  $\hat{Q}$  est environ 8 % plus grande pour la méthode You-Rao que pour MPLSBE. Le prédicteur fondé sur le modèle augmenté à une valeur de  $\hat{Q}$  environ 11 % plus petite que le prédicteur MPLSBE. Dans un certain sens, le modèle augmenté est spécifié moins incorrectement.

Non seulement l'approche d'augmentation donne lieu au calage des prédicteurs sur petits domaines, mais elle réduit aussi le biais au niveau du domaine quand le modèle est spécifié incorrectement. Les tableaux 3 et 4 donnent les propriétés Monte Carlo des prédicteurs pour certains domaines. Dans les tableaux, les biais estimés sont normalisés par  $(\gamma_i m_i^{-1} \sigma^2_{\epsilon})^{0.5}$ , la racine carrée de l'EQM du prédicteur MPLSB avec paramètres connus, et les EQM estimés sont normalisées par  $\gamma_i m_i^{-1} \sigma^2_{\epsilon}$ . Quand le modèle correct (B) est utilisé, le biais Monte Carlo au niveau du domaine approche de zéro. Voir le tableau 3. En outre, l'écart est faible entre l'EQM calculée pour les diverses méthodes, le modèle augmenté produisant une EQM un peu plus grande (moins de 2 %). De nouveau, le tableau 3 montre que le processus de calage a peu d'effet sur l'EQM des prédicteurs au niveau du domaine comparativement au prédicteur MPLSBE.

Tableau 3  
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (B)  
(10 000 échantillons générés par le modèle B)

Etat	Quantité	EMPLSB	PB	ITF	YR	AVG2
1	Biais	0,011	0,012	0,013	0,011	0,011
	EQM	1,100	1,101	1,105	1,104	1,120
2	Biais	0,000	0,001	0,002	0,001	0,001
	EQM	1,072	1,072	1,072	1,076	1,092
14	Biais	-0,001	-0,001	-0,001	-0,001	-0,001
	EQM	1,058	1,058	1,058	1,058	1,074
26	Biais	0,015	0,015	0,015	0,015	0,018
	EQM	1,078	1,077	1,078	1,079	1,092
38	Biais	-0,005	-0,005	-0,005	-0,006	-0,003
	EQM	1,123	1,122	1,122	1,132	1,135
50	Biais	0,012	0,012	0,012	0,009	0,014
	EQM	1,222	1,222	1,222	1,247	1,246

Si nous utilisons un modèle spécifié incorrectement, tel que le modèle (A), le biais de la somme des prédicteurs MPLSBE en tant qu'estimateur du total est négatif dans notre exemple, parce que les Etats pour lesquels le biais est négatif ont une pondération  $\omega_i$  élevée. Voir le tableau 4. Des méthodes d'ajustement, comme PB ou ITF, répartissent le biais entre tous les petits domaines. Donc, l'ajustement réduit le biais négatif des prédicteurs de domaine ayant un grand biais négatif et augmente le biais positif des prédicteurs ayant un grand biais positif. Par conséquent, l'EQM est plus petite pour les grands états et elle est un peu plus grande pour les petits états. Le prédicteur YR produit un biais plus important que le prédicteur ITF. Par ailleurs, les prédicteurs construits avec  $x'_i = (1, \omega_i \hat{\sigma}^2_{\epsilon})$ , c'est-à-dire le modèle augmenté (A), sont nettement supérieurs à ceux construits sous le modèle (A). Le biais au niveau des domaines, que ceux-ci soient grands ou petits, est réduit.

Tableau 4  
Propriétés Monte Carlo des prédicteurs individuels de domaine en utilisant des versions du modèle (A)  
(10 000 échantillons générés par le modèle B).

Etat	Quantité	MPLSBE	PB	ITF	YR	AVG2
1	Biais	-0,597	-0,225	-0,052	-0,473	-0,030
	EQM	1,471	1,165	1,130	1,331	1,139
2	Biais	-0,496	-0,227	-0,170	-0,358	-0,070
	EQM	1,330	1,134	1,115	1,207	1,124
14	Biais	-0,121	0,004	-0,031	0,055	-0,025
	EQM	1,100	1,085	1,086	1,089	1,105
26	Biais	0,057	0,157	0,115	0,249	0,053
	EQM	1,126	1,148	1,136	1,188	1,132
38	Biais	0,380	0,453	0,406	0,601	0,202
	EQM	1,340	1,405	1,361	1,571	1,233
50	Biais	0,922	0,980	0,931	1,178	0,537
	EQM	2,196	2,316	2,215	2,767	1,577

$Y_i$  sur  $x_i$ ,  $k$  est la dimension du vecteur  $x_i$ ,  $\hat{\sigma}_2^2 = m_i^{-1} s_i^2$  et  $s_i^2 = (m_i - 1)^{-1} \sum_{j=1}^m (X_i^j - \bar{Y})^2$  est la variance d'échantillon pour le domaine  $i$ .

Le prédicteur MPLSBE est

$$\hat{Y}_i = x_i' \hat{\beta}_{MCG} + \hat{\gamma}_i (Y_i - x_i' \hat{\beta}_{MCG}) \quad (30)$$

où

$$\hat{\gamma}_i = (\hat{\sigma}_2^2 + \hat{\sigma}_2^{ei})^{-1} \hat{\sigma}_2^2 \quad (31)$$

et

$$\hat{\beta}_{MCG} = \left[ \sum_{i=1}^n (\omega_i^b + \hat{\sigma}_2^{ei})^{-1} x_i x_i' \right]^{-1} \left[ \sum_{i=1}^n (\omega_i^b + \hat{\sigma}_2^{ei})^{-1} x_i Y_i \right] \quad (32)$$

est l'estimateur par les moindres carrés généralisés de  $\beta$ . La

contrainte considérée est  $\sum_{i=1}^{50} \omega_i \gamma_i^* = \sum_{i=1}^{50} \omega_i Y_i$ , où

$$\omega_i = \left( \sum_{i=1}^{50} \text{Pop}_i \right)^{-1} \text{Pop}_i \quad (33)$$

Avec le prédicteur MPLSBE, nous dérivons les prédicteurs PB et ITF en utilisant (13) et (14). Nous dérivons le prédicteur YR en utilisant (16) avec  $\hat{\gamma}_i$  défini en (31) et  $\hat{\beta}_{YR}$  défini en (17). Le prédicteur AUG2 est de la forme (30) avec  $x_i' = (1, \omega_i \hat{\sigma}_2^{ei})$  sous le modèle augmenté (A) et  $x_i' = (1, z_i, \omega_i \hat{\sigma}_2^{ei})$  sous le modèle augmenté (B).

Pour chacun des dix prédicteurs, nous calculons le critère

$$\tilde{Q}(\hat{Y}) = 0,02 \sum_{i=1}^{50} \Phi_i (\hat{Y}_i - Y_i)^2 \quad (34)$$

où  $\Phi_i = (\gamma_i m_i^{-1} \sigma_i^2)^{-1}$ . Notons que  $\Phi_i^{-1}$  est la variance du prédicteur de  $Y_i$  construite à l'aide des paramètres connus.

## 5.2 Résultats de la simulation

L'estimateur de  $\sigma_2^2$  construit sous le modèle (B) a une moyenne Monte Carlo de 1,001 et un écart-type de 0,386. Quand nous utilisons le modèle (A) pour l'estimation, la moyenne Monte Carlo de  $\sigma_2^2$  est 1,720 et l'écart-type est de 0,521. Le carré de la moyenne de  $3z_i$  est 0,636. Par conséquent, la méthode d'estimation avec l'utilisation du modèle (A) intègre une part importante de l'effet fixe de domaine (B) dans l'effet aléatoire. Quand  $\hat{\sigma}_2^2$  est plus grand,  $\hat{\gamma}_i$  est plus grand et une plus forte proportion de  $Y_i$  est utilisée pour construire les prédicteurs, ce qui annule en partie l'effet de la spécification du modèle. Sous le modèle augmenté (A), l'estimateur de  $\sigma_2^2$  est égal à 1,197, c'est-à-dire qu'une part nettement plus faible de l'effet fixe de domaine (B) est intégrée dans l'effet aléatoire.

Le tableau 2 contient certaines statistiques sommaires pour les prédicteurs fondés sur 10 000 échantillons simulés. Le biais empirique dans  $\sum \omega_i (Y_i - \hat{Y}_i)$  est nul pour tous les

prédicteurs quand le modèle (B) ou sa version augmentée est utilisé, parce que le modèle de prédiction concorde avec le modèle de génération des données. L'écart-type simulé de la différence  $\sum \omega_i (Y_i - \hat{Y}_i)$  est égal à 0,022 pour le prédicteur MPLSBE habituel. L'écart-type de  $\sum \omega_i (Y_i - \hat{Y}_i)$  est nul pour les quatre autres prédicteurs, parce que ceux-ci satisfont la contrainte  $\sum_{i=1}^{50} \omega_i \hat{Y}_i = \sum_{i=1}^{50} \omega_i Y_i$ .

**Tableau 2**  
Propriétés Monte Carlo des prédicteurs pour petits domaines (moyenne de 10 000 échantillons générés par le modèle B)

Quantité	MPLSBE	PB	ITF	YR	AUG2
Prédicteur construit sous le modèle (A)					
$\sum \omega_i (Y_i - \hat{Y}_i)$ Moyenne	-0,100	0,000	0,000	0,000	0,000
(E.-T.)	(0,027)	(0,000)	(0,000)	(0,000)	(0,000)
$\tilde{Q}(\hat{Y})$ Moyenne	1,438	1,446	1,419	1,558	1,298
Prédicteur construit sous le modèle (B)					
$\sum \omega_i (Y_i - \hat{Y}_i)$ Moyenne	-0,000	0,000	0,000	0,000	0,000
(E.-T.)	(0,022)	(0,000)	(0,000)	(0,000)	(0,000)
$\tilde{Q}(\hat{Y})$ Moyenne	1,203	1,202	1,202	1,208	1,219

La prédiction fondée sur le modèle (A) ou sa version augmentée est biaisée, parce que le modèle de génération des données (B) contient une fonction de la taille de la population. La moyenne simulée de la différence pondérée  $\sum \omega_i (Y_i - \hat{Y}_i)$  est -0,100, où  $\hat{Y}_i$  est le prédicteur MPLSBE. La valeur de la statistique  $t$  pour le biais pondéré est -3,70. La variance simulée de la moyenne pondérée de la prédiction sur le modèle (A) de  $\sum \omega_i Y_i$  pour les données générées par le modèle (B) est

$$V \left\{ \sum_{i=1}^{50} \omega_i \hat{Y}_i \right\} = 0,060.$$

$$\text{EQM} \left\{ \sum_{i=1}^{50} \omega_i (\hat{Y}_i - Y_i) \right\} = V \left\{ \sum_{i=1}^{50} \omega_i \hat{Y}_i \right\} + \text{Biais}^2 = 0,060 + (-0,100)^2 = 0,070. \quad (35)$$

La variance de  $\sum \omega_i Y_i$  est

$$V \left\{ \sum_{i=1}^{50} \omega_i Y_i \right\} = \sum_{i=1}^{50} \omega_i^2 (\sigma_i^2 + m_i^{-1} \sigma_2^2) = 0,0622. \quad (36)$$

Donc, l'utilisation de  $\sum \omega_i Y_i$  comme estimateur de  $\sum \omega_i Y_i$  est environ 12,5 % plus efficace que celle du prédicteur  $\sum \omega_i \hat{Y}_i$  fondé sur le modèle (A). Étant donné le calage, les quatre prédicteurs (PB, ITF, YR et AUG2) de  $\sum \omega_i Y_i$  ont la même EQM que la moyenne estimée directement  $\sum \omega_i Y_i$ . Le carré du biais représenterait une proportion beaucoup plus importante de l'erreur quadratique moyenne si les petits domaines étaient plus nombreux.



où  $X_i = m_i^{-1} \sum_{j=1}^{m_i} Y_i^{(j)}$  et  $e_i = m_i^{-1} \sum_{j=1}^{m_i} e_{ij}$ . Avec les tailles d'échantillon données au tableau 1, le  $\gamma_i$  défini en (6) est 0,784 pour l'Etat le plus grand (Californie) et 0,304 pour l'Etat le plus petit (Wyoming).

Tableau 1 Populations et tailles d'échantillon pour la

Population	Taille de	Population	Taille de
(en milliers)	(en milliers)	(en milliers)	(en milliers)
Etat	Echantillon	Etat	Echantillon

1	33 640	58	26	4 000	20
2	21 160	46	27	3 610	19
3	19 360	44	28	3 240	18
4	16 000	40	29	3 240	18
5	12 250	35	30	2 890	17
6	12 250	35	31	2 890	17
7	11 560	34	32	2 560	16
8	10 240	32	33	2 560	16
9	8 410	29	34	2 250	15
10	8 410	29	35	1 960	14
11	7 840	28	36	1 690	13
12	7 290	27	37	1 690	13
13	6 250	25	38	1 690	13
14	6 250	25	39	1 210	11
15	5 760	24	40	1 210	11
16	5 760	24	41	1 210	11
17	5 760	24	42	1 210	11
18	5 290	23	43	1 000	10
19	5 290	23	44	810	9
20	5 290	23	45	810	9
21	4 840	22	46	810	9
22	4 410	21	47	640	8
23	4 410	21	48	640	8
24	4 410	21	49	640	8
25	4 000	20	50	490	7

sont connues, nous pouvons construire les variables pour le modèle (23) en utilisant  $\omega_e = (\omega_1 f(u_1), \dots, \omega_n f(u_n))'$  sans estimer  $\sigma_e^2$ . Par exemple, si  $\sigma_e^2 = m_i^{-1} \sigma_e^2$ , le  $\omega_e$  pour le modèle (23) est  $(m_i^{-1} \omega_1, \dots, m_n^{-1} \omega_n)'$ .

#### 4. L'EQM des prédicteurs modifiés

Nous pouvons montrer que tout prédicteur de la forme (15) peut s'écrire

$$\gamma^a = Y - C_a^{-1} B C_a (I_n - I)(Y - X\beta) \quad (24)$$

en posant que  $C_a = A_a^a T$ , où

$$A_a^a = \begin{pmatrix} 1 & -a_{n-1}^{n-1} \\ 0 & I_{n-1}^{n-1} \end{pmatrix}, \quad a_{n-1}^a = (a_2, \dots, a_n)'$$

Par conséquent, nous pouvons utiliser l'estimateur de la variance de  $\gamma^a$  défini en (14) et proposé dans Isaki et coll. (2000) pour estimer l'EQM de tout prédicteur de la forme (15). Souvent, l'EQM d'un prédicteur ajusté est proche de celle du prédicteur avant l'ajustement.

Comme le modèle augmenté (23) possède la propriété d'autocalage, nous pouvons employer, pour estimer l'EQM, la formule utilisée pour les prédicteurs MPLSBE habituels.

### 5. Etude par simulation

#### 5.1 Conditions de simulation

Afin d'étudier les propriétés empiriques des prédicteurs sur petits domaines décrits aux sections 2 et 3, nous utilisons les données conques pour simuler une grande enquête nationale dans laquelle on souhaite produire des estimations par Etat. Le tableau 1 contient les populations approximatives des 50 Etats des Etats-Unis en l'an 2000. Les tailles d'échantillon  $m_i$  données dans le tableau sont approximativement proportionnelles aux racines carrées des populations des Etats. Nous avons produit en tout 10 000 échantillons pour l'étude par simulation. Chacun était composé d'observations produites à l'aide du modèle

$$Y_{ij} = x_i' \beta + b_i + e_{ij}, \quad (25)$$

où  $x_i' = (1, z_i)$ ,  $\beta = (6, 0, 3, 0)$ ,  $z_i = \text{Pop}_{i,0.2}^i / \text{Pop}_{i,0.2}$ , est la moyenne de  $\text{Pop}_{i,0.2}^i$ ,  $b_i \sim N(0, 1)$ , et  $e_{ij} \sim N(0, 16)$ . Les  $b_i$  et les  $e_{ij}$  sont indépendants. Pour les observations au niveau de l'Etat, le modèle devient

$$Y_i = x_i' \beta + b_i + e_i, \quad (26)$$

$c_i = (\sum_{j=1}^{m_i} m_{j,0.5}^{j,0.5})^{-1} m_{i,0.5}^{i,0.5}$  est l'estimateur par les moindres carrés ordinaires du coefficient de régression de

$$\hat{\sigma}_e^2 = \max\{0, 5I^2(\hat{\sigma}_e^2)\}_{0.5}, \hat{\sigma}_{\gamma}^2, \quad (27)$$

Pour étudier les propriétés des cinq prédicteurs, c'est-à-dire MPLSBE, Pfeiffermann-Bernard (PB), Isaki-Tsay-Fuller (ITF), You-Rao (YR) et le modèle augmenté (23) (AUG2), nous nous sommes servis de deux modèles d'estimation. Le premier est un modèle spécifique incorrectement avec  $x_i' = 1$  dans la notation de (26). Nous l'appelons modèle (A). De même, nous démontrons modèle (B) le modèle générateur de données (26) avec  $x_i' = (1, z_i)$ . Suivant la méthode décrite dans Wang et Fuller (2003), l'estimateur de  $\sigma_e^2$  est

$$\hat{\sigma}_e^2 = \sum_{i=1}^I c_i \left[ \frac{50 - k}{50} (Y_i - x_i' \beta_{\text{MCO}})^2 - \hat{\sigma}_{e_i}^2 \right], \quad (28)$$

$$V(\hat{\sigma}_e^2) = \sum_{i=1}^I c_i^2 \left\{ \left[ \frac{50 - k}{50} (Y_i - x_i' \beta_{\text{MCO}})^2 - \hat{\sigma}_{e_i}^2 \right]^2 - \hat{\sigma}_{e_i}^2 \right\}, \quad (29)$$

où les  $\varphi_i, i = 1, \dots, n$  sont un ensemble choisi de poids positifs.

**Théorème 1.** Soit le modèle à effets aléatoires

$$Y_i = x_i' \beta + b_i + e_i, \quad i = 1, \dots, n,$$

où les  $b_i$  sont indépendants et de même loi de moyenne nulle et de variance  $\sigma_b^2$ , les  $e_i$  sont indépendants et de loi de moyenne nulle et de variance  $\sigma_e^2$ , et  $b = (b_1, \dots, b_n)'$  est indépendant de  $e = (e_1, \dots, e_n)'$ . Supposons que  $\sigma_b^2$  et  $\sigma_e^2$  sont connues, et que  $\beta$  est inconnu. Soit  $y_H^i$  le MPLSB de  $y_i$  défini en (5). Soit

$$y_a^i = y_H^i + a_i \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j y_H^j \right), \quad (19)$$

où  $a_i = (\sum_{j=1}^n \varphi_j^{-1} \omega_j^{-1} \varphi_j^{-1})^{-1} \omega_i^{-1}$ , et  $\omega_i$  sont les poids fixes de (7). Alors  $y_a^i = (y_a^1, \dots, y_a^n)'$  est, parmi tous les prédicteurs linéaires sans biais, le seul qui satisfait (7) et minimise le critère (18).

**Preuve :** Voir l'annexe A.

**Remarque 1.** Quand les composantes de la variance sont inconnues, nous les remplaçons dans (6) par des estimateurs appropriés pour obtenir le MPLSB empirique, ou MPLSBE, dénoté par  $\hat{y}_H^i$ . Donc, nous obtenons le prédicteur modifié

$$y_a^i = y_H^i + a_i \left( \sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j \hat{y}_H^j \right). \quad (20)$$

**Remarque 2.** Le critère (18) définit une « fonction de perte », où le choix des poids  $\varphi_i$  dépend du problème étudié. Par exemple, un statisticien peut décider d'attribuer des poids plus élevés aux domaines « plus importants » et des poids plus faibles aux domaines « moins importants ». Souvent,  $\varphi_i$  est une fonction des composantes de la variance. Dans certains cas, on peut choisir  $\varphi_i$  de sorte que les prédicteurs dérivés possèdent certaines propriétés souhaitables. Par exemple,  $\varphi_i = [\widehat{\text{Var}}(Y_j)]^{-1}$  donne les prédicteurs ITF, qui sont les prédicteurs MPLSB (au sens classique) dans le sous-espace qui est orthogonal à  $(\omega_1, \dots, \omega_n)'$  dans l'espace couvert par  $Y$ .

**Remarque 3.** Si  $\varphi_i = \omega_i [\text{cov}(y_H^i, y_j)]^{-1}$ , où  $y_j = \sum_{j=1}^n \omega_j y_H^j$ , nous obtenons le prédicteur (13) dérivé par Pfeffermann et Barnard (1991). Quand  $\varphi_i = [\widehat{\text{Var}}(y_H^j)]^{-1}$ , nous obtenons celui utilisé par Battese, Harter et Fuller (1988).

### 3. Un autre moyen d'imposer la contrainte

Nous avons discuté d'une famille de prédicteurs pour lesquels le total des prédicteurs sur petits domaines est égal

au total des estimations d'enquête directes. L'imposition de la contrainte (7) comporte implicitement la possibilité que le prédicteur sur petits domaines du total présente un biais parce que le modèle (3) est mal spécifié. Dans les applications pratiques, l'erreur de spécification du modèle est une préoccupation valide, car le mécanisme réel de production de  $Y$  est inconnu.

Une erreur de spécification courante survient quand les variables explicatives utilisées dans le modèle ne sont pas les mêmes que celles qui ont produit  $Y$ . Donc, la direction du biais global pourrait ne pas être la même que celle du biais pour un petit domaine particulier. Le cas échéant, les prédicteurs de la forme (15) pourraient accroître le biais pour certains petits domaines comparativement au biais avant ajustement. Mantel et coll. (1993) ont conclu qu'en général, l'effet de l'échantillonnage est une légère amélioration du biais global au prix d'une certaine détérioration des autres mesures d'évaluation.

Puisque le biais n'est pas nul s'il existe une corrélation non nulle entre  $\omega_i$  et  $(Y_j - y_j)$ , on peut le réduire en incluant  $\omega_i$  dans le modèle. Autrement dit, pour un modèle donné, une approche consiste à utiliser le modèle augmenté

$$Y = X_1 \beta + b + e, \quad (21)$$

où  $X_1 = (X, \omega)$  et  $\omega = (\omega_1, \dots, \omega_n)'$  pour obtenir le MPLSB ou MPLSBE. Quand  $\omega$  est dans le modèle, l'ajustement nécessaire pour satisfaire la contrainte (7) sera souvent beaucoup plus faible que celui requis pour le modèle sans  $\omega$ .

Grâce à l'approche du modèle augmenté, nous pouvons aller une étape plus loin et construire des prédicteurs qui satisfont la contrainte (7). En premier lieu, supposons que les variances  $\sigma_e^2$  sont connues. Notons que

$$\sum_{j=1}^n \omega_j Y_j - \sum_{j=1}^n \omega_j y_H^j = \sum_{i=1}^n \omega_i (1 - \gamma_i) (Y_i - x_i' \beta) \quad (22)$$

et  $\omega_i (1 - \gamma_i) \text{Var}(Y_j) = \omega_i \sigma_e^2$ . En appliquant la théorie du modèle linéaire, nous pouvons montrer que le prédicteur construit avec le modèle augmenté

$$Y = X_2 \beta + b + e, \quad (23)$$

où  $X_2 = (X, \omega_e)$  et  $\omega_e = (\omega_1 \sigma_e^2, \dots, \omega_n \sigma_e^2)'$  possède la propriété d'autocalage quand nous utilisons l'estimateur par les moindres carrés généralisés (MCG) de  $\beta$ . Notons que cette approche donne un prédicteur différent de celui de You-Rao (16).

Si les  $\sigma_e^2$  sont inconnues, nous remplaçons  $\sigma_e^2$  dans  $\omega_e$  par son estimateur  $\hat{\sigma}_e^2$ . À condition que les  $\hat{\sigma}_e^2$  dans  $\omega_e$  soient les mêmes que les  $\hat{\sigma}_e^2$  utilisés pour construire  $y_j$ , les prédicteurs ont la propriété d'autocalage. Si les  $\sigma_e^2$  sont de la forme  $\sigma_e^2 f(u_i)$ , où  $\sigma_e^2$  est inconnue, mais que  $u_i$  et  $f(\cdot)$

$$(7) \quad \sum_{i=1}^n \omega_i \hat{y}_i = \sum_n \omega_i Y_i$$

où les  $\omega_i$  sont les poids d'échantillonnage, tels que  $\sum_{i=1}^n \omega_i Y_i$  est un estimateur du total (ou de la moyenne) convergent sous le plan. Un certain nombre de méthodes, souvent appelées « échantillonnage » ou « calage », ont été proposées pour construire des prédicteurs qui satisfont (7). Voir, par exemple, Mantel, Singh et Barreau (1993) et You et Rao (2003).

Afin de passer ce genre de méthodes en revue, représentons par  $\mathbf{y}_H^n = (y_{H1}^n, \dots, y_{Hn}^n)$  le prédicteur MPLSB de  $\mathbf{y} = (y_1, \dots, y_n)$  défini en (5), où

$$(8) \quad \mathbf{y}_H^n = \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{b},$$

et  $\hat{\boldsymbol{\beta}}$  et  $\mathbf{b}$  représentant toute solution de

$$(9) \quad \begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{X} & \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{X} \\ \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{X} & \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{Y} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{Y} \\ \mathbf{X}' \boldsymbol{\Sigma}_1^{-1} \mathbf{Y} \end{bmatrix},$$

$\boldsymbol{\Sigma}_b = \sigma_b^2 \mathbf{I}_n$  et  $\boldsymbol{\Sigma}_e = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ . L'équation (9) est appelée l'équation du modèle mixte. Trouver une solution de cette équation équivaut à trouver une solution au problème de minimisation

$$(10) \quad \min_{\boldsymbol{\beta}, \mathbf{b}} \{ (\mathbf{X} - \mathbf{X} \boldsymbol{\beta} - \mathbf{b})' \boldsymbol{\Sigma}_e^{-1} (\mathbf{X} - \mathbf{X} \boldsymbol{\beta} - \mathbf{b}) + \mathbf{b}' \boldsymbol{\Sigma}_b^{-1} \mathbf{b} \},$$

Pfeffermann et Barnard (1991) ont proposé le prédicteur modifié

$$(11) \quad \mathbf{y}_{PB} = \mathbf{X} \hat{\boldsymbol{\beta}}_{PB} + \mathbf{b}_{PB},$$

où  $\hat{\boldsymbol{\beta}}_{PB}$  et  $\mathbf{b}_{PB}$  représentent toute solution du problème de minimisation (10) avec  $\hat{\boldsymbol{\beta}}$  et  $\mathbf{b}$  soumis à la contrainte

$$(12) \quad \sum_{i=1}^n \omega_i (\mathbf{x}_i' \hat{\boldsymbol{\beta}} + b_i) = \sum_n \omega_i Y_i.$$

Cela mène au prédicteur

$$(13) \quad \mathbf{y}_{PB} = \mathbf{y}_H^n + [\text{Var}(\hat{\mathbf{y}})]^{-1} \text{cov}(\hat{\mathbf{y}}_H^n, \hat{\mathbf{y}}) \left[ \sum_{i=1}^n \omega_i Y_i - \hat{\mathbf{y}} \right],$$

où  $\hat{\mathbf{y}} = \sum_{i=1}^n \omega_i \hat{y}_i$ ,  $\text{cov}(\hat{\mathbf{y}}_H^n, \hat{\mathbf{y}}) = \omega_i Y_i \sigma_i^2 + \sum_{i=1}^n \omega_i \text{cov}(\hat{y}_H^n, \hat{y})$ , et  $\text{Var}(\hat{\mathbf{y}}) = \sum_{i=1}^n \omega_i (1 - \gamma_i) \mathbf{x}_i' \mathbf{x}_i + \text{Var}(\hat{\mathbf{y}})$ .

Isaki, Tsay et Fuller (2000) ont imposé la contrainte par une méthode qui, approximativement, consiste à construire les meilleurs prédicteurs de  $n - 1$  quantités qui ne sont pas corrélées avec  $\sum_{i=1}^n \omega_i Y_i$ . Après certaines opérations matricielles, le prédicteur d'Isaki-Tsay-Fuller (ITF) peut être décrit sous la forme

$$\mathbf{y}_{ITF}^n = \mathbf{y}_H^n +$$

$$(14) \quad \left[ \sum_{i=1}^n \omega_i^2 \text{Var}(Y_i) \right]^{-1} \left[ \omega_i \text{Var}(Y_i) \left( \sum_{i=1}^n \omega_i Y_i - \sum_{i=1}^n \omega_i \hat{y}_H^n \right) \right],$$

où  $\text{Var}(Y_i)$  est un estimateur de  $\sigma_i^2 + \sigma_e^2$ . Soulignons que le prédicteur de Pfeffermann-Barnard (PB) (13) et le prédicteur ITF (14) sont de la forme

$$(15) \quad \mathbf{y}_a^n = \mathbf{y}_i + a_i \left( \sum_{i=1}^n \omega_i Y_i - \sum_n \omega_i \hat{y}_i \right),$$

où  $\sum_{i=1}^n \omega_i a_i = 1$ . Autrement dit, nous pouvons considérer qu'imposer la contrainte (7) revient à un problème d'ajustement. Pour qu'un prédicteur ajusté  $\mathbf{y}_a^n$  satisfasse (7), nous allouons la différence  $\sum_{i=1}^n \omega_i Y_i - \sum_{i=1}^n \omega_i \hat{y}_i$  aux prédicteurs sur petits domaines  $\hat{y}_i$  en utilisant  $a_i$ .

A l'aide du modèle au niveau des unités, You et Rao (2002) ont proposé un estimateur de  $\boldsymbol{\beta}$  tel que les prédicteurs résultants satisfont (7). Ils ont dit que ces prédicteurs étaient autocalés. En appliquant leur méthode au modèle au niveau du domaine (3), nous obtenons

$$(16) \quad \mathbf{y}_{YR}^i = \hat{\gamma}_i Y_i + (1 - \hat{\gamma}_i) \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{YR}^i,$$

où

$$(17) \quad \hat{\boldsymbol{\beta}}_{YR} = \left[ \sum_{i=1}^n \omega_i (1 - \hat{\gamma}_i) \mathbf{x}_i' \mathbf{x}_i \right]^{-1} \left[ \sum_{i=1}^n \omega_i (1 - \hat{\gamma}_i) \mathbf{x}_i' Y_i \right]$$

Tout prédicteur qui a la propriété d'autocalage, tel que le prédicteur (16) de You et Rao (YR), est un prédicteur de la forme (15), puisque la différence  $\sum_{i=1}^n \omega_i Y_i - \sum_{i=1}^n \omega_i \hat{y}_i$  est égale à zéro.

À la section 2, nous dérivons le « meilleur » prédicteur de la forme (15). Les résultats conduisent à une vision unifiée de plusieurs prédicteurs fondés sur le MPLSB. À la section 3, nous proposons une autre approche qui possède la propriété d'autocalage. À la section 4, nous décrivons brièvement l'erreur quadratique moyenne (EQM) et à la section 5, nous recourons à des études par simulation pour comparer les prédicteurs. À la section 6, nous présentons nos conclusions et une discussion.

## 2. « Meilleur » prédicteur linéaire sans biais sous une contrainte

Afin de trouver le « meilleur » prédicteur linéaire sans biais de  $\mathbf{y}$  qui satisfait la contrainte (7), nous commençons par supposer que les paramètres pour les composantes de la variance sont connus. Selon le lemme 1 de Pfeffermann et Barnard (1991), il est impossible de comparer, composante par composante, des prédicteurs qui satisfont la contrainte (7) pour trouver le meilleur d'entre eux. Par conséquent, un certain critère global est requis. L'un des critères naturels est

$$(18) \quad \tilde{O}(\mathbf{y}_a^n) = \sum_{i=1}^n \phi_i E(G_i^n - Y_i)^2,$$



## Estimation pour petits domaines sous une contrainte

Junyuan Wang, Wayne A. Fuller et Yongming Qu<sup>1</sup>

### Résumé

La prédiction sur petits domaines fondée sur des effets aléatoires, appelée (MPLSBE), est une méthode de construction d'estimations pour de petites régions géographiques ou de petites sous-populations en utilisant les données d'enquête existantes. Souvent, le total des prédicteurs sur petits domaines est forcé d'être égal à l'estimation par sondage directe et ces prédicteurs sont alors dits calés. Nous passons en revue plusieurs prédicteurs calés et présentons un critère qui unit leur calcul. Nous dérivons celui qui est l'unique meilleur prédicteur linéaire sans biais sous ce critère et discutons de l'erreur quadratique moyenne des prédicteurs calés. L'imposition de la contrainte comporte implicitement la possibilité que le modèle de petit domaine soit spécifié incorrectement et que les prédicteurs présentent un biais. Nous étudions des modèles augmentés contenant une variable explicative supplémentaire pour lesquels les prédicteurs sur petits domaines ordinaires consistent à utiliser un modèle augmenté contenant une variable auxiliaire supplémentaire qui est fonction de la taille du domaine. Dans les simulations, les prédicteurs fondés sur le modèle augmenté ont une EQM plus petite que MPLSBE quand le modèle incorrect est utilisé pour la prédiction. De surcroît, l'EQM augmente très légèrement comparativement à celle de MPLSBE si la variable auxiliaire est ajoutée au modèle correct.

Mots clés : Modèle des composantes de la variance; meilleure prédiction linéaire sans biais; calage; convergence sous le plan; modèles linéaires mixtes.

### 1. Introduction

Dans certaines situations, il est souhaitable d'établir d'après les données d'enquête existantes des estimateurs fiables pour de petites régions géographiques ou de petites sous-populations. Cependant, les tailles des échantillons de domaine peuvent être telles que les estimateurs habituels produisent des erreurs-types inacceptablement grandes et il est alors raisonnable d'utiliser un estimateur fondé sur un modèle. Voir Rao (2003) pour une discussion complète de l'estimation pour petits domaines.

Un modèle pour l'estimation pour petits domaines peut s'écrire

$$y_i = x_i' \beta + b_i \quad (1)$$

$$x_i = y_i + e_i, \quad i = 1, \dots, n, \quad (2)$$

où les  $y_i$  sont les moyennes inobservables de petits domaines, les  $x_i$  sont les estimateurs d'après les données d'enquête observables, les  $x_i'$  sont des vecteurs connus,  $\beta$  est le vecteur des paramètres de régression, les  $b_i$  sont des variables aléatoires indépendantes et de même loi telles que  $E(b_i) = 0$  et  $V(b_i) = \sigma_b^2$  et les  $e_i$  sont les erreurs d'échantillonnage telles que  $E(e_i | y_i) = 0$  et  $V(e_i | y_i) = \sigma_e^2$ . En combinant (1) et (2), nous obtenons

$$y_i = x_i' \beta + b_i + e_i, \quad i = 1, \dots, n, \quad (3)$$

qui est un cas particulier du modèle linéaire mixte.

En supposant que les composantes de la variance  $\sigma_b^2$  et  $\sigma_e^2$  sont connues, le meilleur estimateur linéaire sans biais de  $\beta$  est

$$\hat{\beta} = [X' \Sigma^{-1} X]^{-1} X' \Sigma^{-1} Y \quad (4)$$

$$= \left[ \sum_{i=1}^n (\sigma_b^2 + \sigma_e^2)^{-1} x_i x_i' \right]^{-1} \left[ \sum_{i=1}^n (\sigma_b^2 + \sigma_e^2)^{-1} x_i y_i \right],$$

où  $X' = (x_1', \dots, x_n')$ ,  $Y' = (y_1, \dots, y_n)$  et  $\Sigma = \text{Var}(Y) = \text{diag}(\sigma_b^2 + \sigma_e^2, \dots, \sigma_b^2 + \sigma_e^2)$ . En outre, le meilleur prédicteur linéaire sans biais (MPLSBE) de  $y_i$  est

$$\hat{y}_H = x_i' \hat{\beta} + y_i(Y'Y)^{-1} - x_i' \hat{\beta}, \quad (5)$$

$$y_i = (\sigma_b^2 + \sigma_e^2)^{-1} \sigma_b^2. \quad (6)$$

Voir Henderson (1963) et Rao (2003). Quand les composantes de la variance sont inconnues, nous les remplaçons dans (4) et (6) par des estimateurs pour obtenir  $\hat{y}_H$ , c'est-à-dire le MPLSBE empirique ou MPLSBE.

Souvent, l'estimateur direct du total de tous les domaines étudiés est considéré comme étant d'une précision adéquate. Si cela est le cas, le praticien pourrait choisir d'utiliser l'estimateur du total convergent sous le plan et exiger que la somme pondérée des prédicteurs sur petits domaines soit égale à cet estimateur. Donc, il est désirable que les prédicteurs sur petits domaines  $\hat{y}_i$  satisfassent



You, Y., et Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. et Gambino, J. (2000). Hierarchical Bayes estimation of unemployment rates for sub-provincial regions using cross-sectional and time series data. *American Statistical Association 2000 Proceedings of the Section on Government Statistics and Section on Social Statistics*, 160-165.

You, Y., Rao, J.N.K. et Gambino, J. (2003). Estimation du taux de chômage fondée sur un modèle pour l'Enquête sur la population active du Canada : une approche bayésienne hiérarchique. *Techniques d'enquête*, 29, 27-36.



- Pour  $i = 1, \dots, m$ ,  

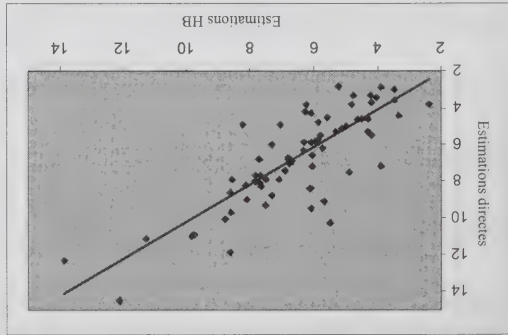
$$n_{i,T}^{IT} \gamma_i \beta_i \sigma_v^2 \sigma_\varepsilon^2 n_{i,T-1}^{\varepsilon^2} \theta \sim N \left( \left( \frac{1}{1} + \frac{\sigma_v^2}{\sigma_\varepsilon^2} \right)^{-1} \left( \log(\theta_{i,T}) - x_{i,T}^T \beta + \frac{\sigma_v^2}{n_{i,T-1}^{\varepsilon^2}} + \frac{\sigma_\varepsilon^2}{n_{i,T-1}^{\varepsilon^2}} \right), \left( \frac{1}{1} + \frac{\sigma_v^2}{\sigma_\varepsilon^2} \right)^{-1} \right);$$
 • Pour  $i = 1, \dots, m$ ,  

$$\theta_i | \gamma_i, \beta_i, \sigma_v^2, \sigma_\varepsilon^2, n \propto \left\{ \sum_{j=1}^n \exp \left\{ -\frac{1}{2} (y_j - \theta_i)' \sum_{l=1}^n (y_l - \theta_i) \right\} \right\} \times \exp \left\{ -\frac{1}{2\sigma_v^2} \sum_{i=1}^n (\log(\theta_{i,T}) - x_{i,T}^T \beta - n_{i,T}^{\varepsilon^2}) \right\} \left( \prod_{i=1}^n \frac{\theta_i}{1} \right)^{\frac{1}{2}}.$$
- Bayarri, M.J., et Berger, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Brown, G., Chambers, R., Headay, P., et Heasman, D. (2001). Evaluation of small area estimation methods - An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency: A Methodological Perspective*. CD-ROM.
- Chib, S., et Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.
- Datta, G.S., Day, B., et Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sanhyā*, 60, 344-362.
- Datta, G.S., Lahiri, P., Maiti, T. et Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Datta, G.S., Lahiri, P. et Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.
- Dick, P. (1995). Modélisation du sous-dénombrement net dans le recensement du Canada de 1991. *Techniques d'enquête*, 21, 51-61.
- Statistique Canada, N° 12-001-X au catalogue
- Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.
- Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. et Lindsey, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistique Canada, Catalogue No. 71-526.
- Gelfand, A.E., et Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.
- Gelman, A., Carlin, J.B., Stern, H.S., et Rubin, D.B. (1995). *Bayesian Data Analysis*. Londres : Chapman and Hall.
- Gelman, A., et Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Ghosh, M., Nanjia, N. et Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.
- Laud, P., et Ibrahim, J. (1995). Predictive model selection. *Journal of Royal Statistical Society, Series B*, 57, 247-262.
- Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.
- Rao, J.N.K., et Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York : John Wiley & Sons, Inc.
- Singh, A.C., Folsom, R.E., Jr. et Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical methods Conference proceedings*. Washington, D.C., [www.fcsnm.gov](http://www.fcsnm.gov)
- Singh, A., You, Y. et Mantel, H. (2005). Use of generalized design effects for variance function modeling in small area estimation from survey data. Présentation au 2005 Statistical Society of Canada Annual Meeting, Regina, SK.
- Sinharay, S., et Stern, H.S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.
- Wang, J., et Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- You, Y., et Chapman, B. (2006). Estimation pour petits domaines au moyen de modèles régionaux et d'estimations des variances d'échantillonnage. *Techniques d'enquête*, 32, 107-114.
- You, Y., Chen, E. et Gambino, G. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics*. Alexandria, VA : American Statistical Association. 3883-3888.
- You, Y., et Dick, P. (2004). Hierarchical Bayes small area inference to the 2001 census undercoverage estimation. *2004 Proceedings of the American Statistical Association, Section on Government Statistics*. Alexandria, VA : American Statistical Association. 1836-1840.

d'établir un modèle fiable et facile à mettre en œuvre de l'estimation des taux de chômage fondés sur un modèle ajusté aux données de l'EPA pour les petits domaines. Nous prévoyons produire les estimations fondées sur le modèle pour une période relativement longue, par exemple 24 mois allant de 2004 à 2005. Nous comparerons ces estimations aux estimations directes pour les 24 mois, particulièrement pour les grandes RMR, afin d'étudier les effets de lissage du modèle proposé. Les estimations fondées sur le modèle devraient suivre le profil des estimations directes d'après l'EPA pour les grandes RMR, ce qui indiquerait que le lissage des effets de la série temporelle sont raisonnables. Le but est de vérifier la robustesse des estimations fondées sur le modèle proposé au fil du temps.

## 6. Conclusion et futurs travaux

Figure 4 Diagramme de dispersion avec droite de régression



Nous avons passé en revue certains modèles pour petits domaines, y compris le modèle de Fay-Herriot et le modèle transversal et chronologique de You et coll. (2003). Étant donné les travaux antérieurs, nous avons proposé un modèle non linéaire transversal et chronologique intégré en vue d'obtenir des estimations fondées sur un modèle des taux de chômage pour les RMR/CU au Canada en utilisant les données de l'EPA. Le modèle proposé surmonte les limites constatées lors des travaux antérieurs. En particulier, nous pouvons modéliser la variance d'échantillonnage sous forme d'une fonction de la moyenne de petits domaines en émettant l'hypothèse d'un coefficient de variation constant pour un domaine donné ou d'un effet de plan constant pour un domaine donné. Notre analyse des données révèle que le modèle proposé est relativement bien ajusté aux données. Les estimations hiérarchiques bayésiennes fondées sur le modèle améliorèrent significativement les estimations directes par sondage en ce qui concerne la réduction du coefficient de variation, surtout pour les CU dont la population est

Nous prévoyons adopter l'approche de modélisation de rechange pour la variance d'échantillonnage. Récemment, You et Dick (2004), ainsi que You et Chapman (2006) ont utilisé l'approche HB pour modéliser directement la variance d'échantillonnage, sans spécifier la forme de cette dernière sous le cadre du modèle de Fay-Herriot. Le modèle tient compte automatiquement de la variabilité de l'estimation des variances d'échantillonnage. En particulier, You et Dick (2004) ont appliqué le modèle au problème de l'estimation du sous-dénominateur au recensement et ont obtenu des estimations HB efficaces de ce sous-dénominateur intéressant pour divers petits domaines au Canada. Il serait intéressant d'adapter la même idée au modèle transversal et chronologique et de comparer les résultats à ceux des présents travaux. L'objectif de la comparaison est

- Pour  $i = 1, \dots, m$ , et  $2 \leq i \leq T-1$ ,  

$$n_{it} | Y, \beta, \sigma_v^2, \sigma_e^2, n_{it-1}^2, n_{it+1}^2, \theta \sim \left( \frac{\sigma_v^2}{1} + \frac{\sigma_e^2}{2} \right)^{-1};$$
- Pour  $i = 1, \dots, m$ ,  

$$n_{it} | Y, \beta, \sigma_v^2, \sigma_e^2, n_{it-1}^2, n_{it+1}^2, \theta \sim \left( \frac{\sigma_v^2}{1} + \frac{\sigma_e^2}{2} \right)^{-1};$$
- $\sigma_e^2 | Y, \beta, \sigma_v^2, n, \theta \sim \text{IG} \left( a_1 + mT/2, b_1 + \sum_{t=1}^T \sum_{i=1}^m (n_{it}^2 - n_{it-1}^2) / 2 \right);$
- $\sigma_v^2 | Y, \beta, \sigma_e^2, n, \theta \sim \text{IG} \left( a_2 + m(T-1)/2, b_2 + \sum_{t=1}^T \sum_{i=1}^m (n_{it}^2 - n_{it-1}^2) / 2 \right);$
- $\beta | Y, \sigma_v^2, \sigma_e^2, n, \theta \sim N((X'X)^{-1}X'(\log(\theta) - n), \sigma_v^2, \sigma_e^2(X'X)^{-1});$

il suit :

Nous présentons ci-après les lois conditionnelles complètes pour l'échantillonnage de Gibbs sous le modèle HB proposé. Soit  $Y = (Y_1^1, \dots, Y_1^m)$ ,  $X = (X_1^1, \dots, X_1^m)$ ,  $\theta = (\theta_1^1, \dots, \theta_1^m)$  et  $n = (n_1^1, \dots, n_1^m)$ , avec  $Y_i^t = (Y_{i1}^t, \dots, Y_{it}^t)$ ,  $X_i^t = (X_{i1}^t, \dots, X_{it}^t)$ ,  $\theta_i^t = (\theta_{i1}^t, \dots, \theta_{it}^t)$ , et  $n_i^t = (n_{i1}^t, \dots, n_{it}^t)$ . Nous obtenons les lois conditionnelles complètes comme

$$p = P(D(y^{\text{resp}}, \theta) \geq D(y^{\text{obs}}, \theta) | y^{\text{obs}}).$$

Il s'agit d'une extension naturelle de la valeur  $p$  habituelle dans un contexte bayésien. Si un modèle est ajusté aux données observées, les deux valeurs de la mesure de divergence sont semblables. Autrement dit, si le modèle est ajusté correctement aux données observées,  $D(y^{\text{obs}}, \theta | y^{\text{obs}})$  devrait se situer près de la partie centrale de l'histogramme des valeurs  $D(y^{\text{resp}}, \theta | y^{\text{obs}})$  si  $y^{\text{resp}}$  est généré de façon répétée à partir de la loi prédictive a posteriori. Conséquemment, la valeur  $p$  prédictive a posteriori devrait s'approcher de 0,5 si le modèle est bien ajusté aux données. Les valeurs  $p$  extrêmes (proches de 0 ou de 1) impliquent que l'ajustement n'est pas bon. Nous pouvons estimer la valeur  $p$  prédictive a posteriori de la façon suivante. Soit  $\theta^*$  un tirage à partir de la loi a posteriori  $f(\theta | y^{\text{obs}})$ , et soit  $y^{\text{resp},*}$  un tirage à partir de  $f(y^{\text{resp}} | \theta^*)$ . Alors, marginalement,  $y^{\text{resp},*}$  est un échantillon provenant de la loi prédictive a posteriori  $f(y^{\text{resp}} | y^{\text{obs}})$ . Il est relativement facile de calculer la valeur  $p$  en utilisant les valeurs simulées de  $\theta^*$  provenant de l'échantillonnage de Gibbs. Pour chaque valeur simulée  $\theta^*$ , nous pouvons simuler  $y^{\text{resp},*}$  à l'aide du modèle et calculer  $D(y^{\text{resp},*}, \theta^*)$  et  $D(y^{\text{obs}}, \theta^*)$ . Alors, la valeur  $p$  est estimée par la proportion de fois que  $D(y^{\text{resp},*}, \theta^*)$  excède  $D(y^{\text{obs}}, \theta^*)$ .

Pour le modèle HB proposé, la mesure de divergence utilisée pour l'ajustement global est donnée par

$$d(y, \theta) = \sum_{i=1}^n (y_i - \theta_i)' \Sigma_i^{-1} (y_i - \theta_i).$$

Cette mesure a été utilisée par Datta et coll. (1999) et par You et coll. (2003). Nous avons calculé la valeur  $p$  en combinant les valeurs

simulées de  $\theta^*$  et  $y^{\text{resp},*}$  provenant des dix exécutions parallèles. Nous avons obtenu une valeur  $p$  moyenne

estimée d'environ 0,38. Donc, rien n'indique que l'ajustement global du modèle est insuffisant.

Certains ont critiqué la vérification de l'ajustement du modèle à l'aide de la valeur  $p$  prédictive a posteriori, la

jugeant modérée à cause du double usage des données observées. Le double utilisation des données peut induire un

comportement non naturel, comme l'ont démontré Bayarri et Berger (2000). Ces auteurs ont proposé deux mesures de

la valeur  $p$  de rechange pour la vérification du modèle qu'ils ont nommées la valeur  $p$  prédictive a posteriori partielle et la

valeur  $p$  prédictive conditionnelle. Cependant, leurs méthodes sont plus difficiles à appliquer et à interpréter (Rao

2002; Sinharay et Stern 2003). Comme le soulignent Sinharay et Stern (2003), la valeur  $p$  prédictive a posteriori

est particulièrement utile si nous voyons dans le modèle courant un point final plausible auquel des modifications ne

seront apportées que si un manque considérable d'ajustement est constaté.

Pour comparer le modèle proposé au modèle de You et coll. (2003), nous avons calculé la mesure de divergence de Laud et Ibrahim (1995) fondée sur la loi

### 5.3 Diagnostic du biais par analyse de régression

Afin d'évaluer le biais que pourrait introduire le modèle,

nous utilisons une méthode simple d'analyse de régression par les moindres carrés ordinaires pour les estimations

directes d'après l'EPA ainsi que pour les estimations fondées sur le modèle HB. La méthode de régression est

proposée par Brown, Chamber, Headly et Heasman (2001). Si les estimations fondées sur le modèle sont proches des

taux réels de chômage, les estimateurs directs d'après l'EPA les valeurs prévues correspondent aux valeurs des esti-

mations fondées sur le modèle. Nous représentons graphiquement les estimations HB fondées sur le modèle en

abscisse  $X$  et les estimations directes d'après l'EPA en ordonnée  $Y$  et nous voyons dans quelle mesure la droite de

régression s'approche de  $Y = X$ . En termes de régression, fondamentalement, nous ajustons le modèle de régression,

$Y = \alpha X$  aux données et estimons le coefficient  $\alpha$ . Des estimations fondées sur le modèle dont le biais est faible

devrait donner une valeur de  $\alpha$  proche de 1. Pour les données de juin 2005, soit  $X$  les estimations directes du

taux de chômage et  $X$  les estimations HB fondées sur le modèle. Nous obtenons une estimation de la valeur de  $\alpha$  de

1,0207 avec une erreur-type de 0,0281. La figure 4 donne le diagramme de dispersion avec la droite de régression

ajustée.

Les résultats de la régression ne révèlent aucun écart significatif par rapport à  $Y = X$ . Par conséquent, nous

concluons que les estimations fondées sur le modèle calculées d'après le modèle proposé concordent avec les

estimations directes d'après l'EPA sans inclusion d'un biais éventuel supplémentaire. Les résultats pourraient également

indiquer qu'il n'existe aucune preuve d'un biais dû à une erreur de spécification possible du modèle.



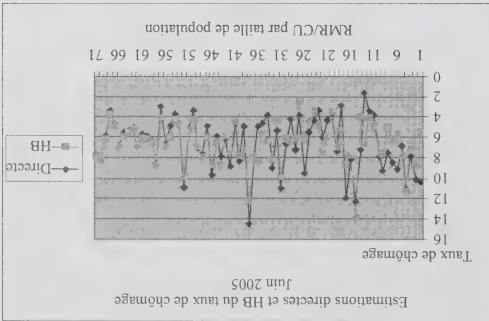
Un CV (Mitrarnicht) ne possède pas de données sur l'assurance-emploi. Donc, nous considérons dans le modèle  $m = 71$  RMR/CU. Dans chaque domaine, nous considérons six estimations mensuelles consécutives  $y^u$ , allant de janvier 2005 à juin 2005, de sorte que  $T = 6$ . Pour les données recueillies de janvier à juin 2005, le taux de chômage moyen global (sur les 71 RMR/CU et les six mois) est de 0,076 et le taux moyen global de bénéficiaires de l'assurance-emploi est de 0,059. Pour le modèle pour petits domaines proposé, le paramètre d'intérêt  $\theta_{IT}$  est le taux de chômage réel pour le domaine  $i$  en juin 2005, où  $i = 1, \dots, 71$ . Pour appliquer l'échantillonnage de Gibbs, nous avons utilisés dix exécutions parallèles comptant chacune 2 000 itérations. Nous avons supprimé les 1 000 premières itérations à titre de période de « rodage ». Les hyperparamètres des composantes de variance incluses dans le modèle sont fixés à 0,0001 pour refléter nos connaissances vagues au sujet de  $\sigma_0^2$  et de  $\sigma_0^2$ .

Nous présentons maintenant les estimations a posteriori des taux de chômage sous le modèle HB intégré proposé décrit à la section 4.1 en utilisant la méthode d'échantillonnage de Gibbs. La figure 2 donne les estimations directes d'après l'EPA et les estimations fondées sur le modèle HB

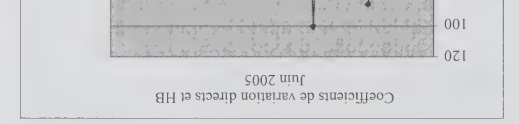
Canada. Ces 71 RMR/CU figurent par ordre de taille de population, en commençant à gauche par le plus petit CU (Dawson Creek, C.-B.) et en terminant à droite par la plus grande RMR (Toronto, Ont.). Dans le cas des estimations ponctuelles, les estimations HB correspondent à un lissage modéré des estimations directes d'après l'EPA. Pour les RMR ayant les populations les plus grandes et, par conséquent, les tailles d'échantillon les plus grandes, les estimations directes et les estimations HB sont très proches, comme il fallait s'y attendre, particulièrement pour Toronto, Montréal et Vancouver; pour les CU plus petits, l'écart entre les estimations directes et HB est important pour certaines régions.

La figure 3 donne les CV des estimations. Le CV de l'estimation HB correspond au ratio de la racine carrée de la variance a posteriori à la moyenne a posteriori. Il est manifeste, si l'on examine la figure 3, que les CV des estimations directes sont très grands, particulièrement pour les CU dont les CV sont très grands et instables. Comparativement aux estimations directes, les CV des estimations HB sont très petits et stables. Le gain d'efficacité obtenu grâce aux estimations HB est évident, particulièrement pour les CU dont la taille de population est faible. Plus précisément, nous avons calculé la réduction en pourcentage des CV pour les estimateurs HB en nous basant sur les données de juin 2005. La réduction en pourcentage du CV est calculée comme la différence entre le CV direct et le CV HB par rapport au CV direct. La réduction moyenne du CV

Figure 2 Comparaison des estimations directes et HB



d'échantillon est faible.



5.2 Ajustement du modèle au moyen de la loi prédictive a posteriori

Pour vérifier l'ajustement global du modèle proposé, nous utilisons la méthode de la loi prédictive a posteriori. Soit  $y^{rep}$  l'observation répétée sous le modèle. La loi prédictive a posteriori de  $y^{rep}$ , sachant les données observées  $y^{obs}$  est définie comme étant

$$f(y^{rep} | y^{obs}) = \int f(y^{rep} | \theta) f(\theta | y^{obs}) d\theta.$$

Selon cette approche, nous pouvons définir une mesure de divergence  $D(y, \theta)$  qui dépend des données  $y$  et du paramètre  $\theta$ , et comparer la valeur observée  $D(y^{obs}, \theta | y^{obs})$  à la loi prédictive a posteriori de  $D(y^{rep}, \theta | y^{obs})$ , tout écart significatif indiquant une défaillance du modèle. Meng (1994), ainsi que Gelman, Carlin, Stern et Rubin (1995) ont proposé la valeur  $p$  pré-

où

$$h(\theta_i) = \left| \sum_i \theta_i \right|^{-1} \exp \left\{ -\frac{2}{1} (y_i - \theta_i)' \sum_{i=1}^T (y_i - \theta_i) \right\}$$

et

$$f(\theta_i) =$$

$$\exp \left\{ -\frac{1}{2\sigma_y^2} (\log(\theta_i) - x_i' \beta - n_i) (\log(\theta_i) - x_i' \beta - n_i) \right\} \cdot \left( \prod_{i=1}^T \frac{\theta_i}{1} \right)^{-1}$$

Pour mettre à jour  $\theta_i$ , nous procédons comme il suit :

1. Pour  $t=1, \dots, T$ , tirer  $\theta_{(k+1)}'' \sim \log N(x_t' \beta''_{(k+1)} + n_{(k+1)}'', \sigma_{2(k+1)}^2)$ , ce qui nous donne alors  $\theta_{(k+1)}'' = (\theta_{1(k+1)}'', \dots, \theta_{T(k+1)}'')$ ;

2. Calculer la probabilité de rejet

$$\alpha(\theta_{(k)}^i, \theta_{(k+1)}^i) = \min \left\{ h(\theta_{(k+1)}^i), 1 \right\}.$$

3. Générer  $\lambda \sim \text{Uniforme}(0,1)$ , si  $\lambda < \alpha(\theta_{(k)}^i, \theta_{(k+1)}^i)$ , alors accepter  $\theta_{(k+1)}^i$ , sinon, rejeter  $\theta_{(k+1)}^i$  et fixer  $\theta_{(k+1)}^i = \theta_{(k)}^i$ .

Pour appliquer l'échantillonnage de Gibbs, nous suivons les recommandations de Gelman et Rubin (1992) et nous exécutons indépendamment ( $L > 2$ ) chaînes parallèles, chacune de longueur  $2d$ . Nous supprimons les  $d$  premières itérations de chaque chaîne. La surveillance de la convergence est fondée sur le facteur de réduction d'échelle possible proposé dans Gelman et Rubin (1992) et adopté par You et coll. (2003) pour estimer  $\theta^{IT}$ . Une description détaillée figure dans You et coll. (2003). Les estimations de la moyenne  $a_{posterior} E(\theta^{IT} | y)$  et de la variance à posteriori  $V(\theta^{IT} | y)$  sont obtenues en se basant sur les échantillons créés à partir de l'échantillonneur de Gibbs.

## 5. Application aux données de l'EPA

### 5.1 Estimation

Nous utilisons les estimations du taux de chômage fondées sur les données de l'EPA recueillies de janvier à juin 2005,  $y^n$ , dans notre analyse de données. En plus des estimations directes  $y^n$ , et des matrices de covariances d'échantillonnage utilisées dans les modèles pour petits domaines, nous avons besoin de variables auxiliaires administratives dans nos modèles. Pour l'estimation du taux de chômage, nous utilisons comme données auxiliaires les taux de bénéficiaires de l'assurance-emploi (a-e.) au niveau local. Le taux de bénéficiaires est calculé sous forme du ratio du nombre de personnes faisant la demande de prestations d'assurance-emploi sur le nombre de personnes dans la population active. Le Canada compte 72 RMR/CU.

### 4.2 Inférence sous échantillonnage de Gibbs

La méthode d'échantillonnage de Gibbs est une méthode itérative d'échantillonnage Monte Carlo par chaînes de Markov utilisée pour simuler le tirage d'échantillons à partir d'une loi conjointe de variables aléatoires en procédant à un échantillonnage à partir de densités sous un espace de dimension réduite pour faire des inférences au sujet des lois conjointe et marginale (Gelfand et Smith 1990). Cette méthode est appliquée principalement à l'inférence dans un cadre bayésien où l'on s'intéresse à la loi a posteriori des paramètres. Supposons que la densité conditionnelle de  $y_i | \theta$  est  $f(y_i | \theta)$  pour  $i=1, \dots, n$  et que l'information a priori au sujet de  $\theta = (\theta_1, \dots, \theta_k)'$  est résumée par une densité a priori  $\pi(\theta)$ . Soit  $\pi(\theta | y)$  la densité a posteriori de  $\theta$  sachant les données  $y = (y_1, \dots, y_n)'$ . En pratique, il pourrait être difficile de tirer directement des échantillons à partir de  $\pi(\theta | y)$ , à cause de l'intégration en grande dimension par rapport à  $\theta$ . Cependant, nous pouvons utiliser l'échantillonneur de Gibbs pour construire une chaîne de Markov  $\{\theta^{(s)} = (\theta_1^{(s)}, \dots, \theta_k^{(s)})' \}$  avec  $\pi(\theta | y)$  comme limite. À titre d'illustration, posons que  $\theta = (\theta_1, \theta_2)'$ . En partant d'un ensemble initial de valeurs  $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$ , nous générons  $\theta^{(s)} = (\theta_1^{(s)}, \theta_2^{(s)})'$  par échantillonnage de  $\theta^{(s)} | \theta^{(s-1)}, y$  à partir de  $\pi(\theta_1 | \theta_2^{(s-1)}, y)$  et  $\theta_2^{(s)}$  à partir de  $\pi(\theta_2 | \theta_1^{(s-1)}, y)$ . Sous certaines conditions de régularité,  $\theta^{(s)} = (\theta_1^{(s)}, \theta_2^{(s)})'$  converge en loi vers  $\pi(\theta | y)$  quand  $g \rightarrow \infty$ . Si la valeur de  $g$  est grande, l'inférence marginale au sujet de  $\pi(\theta_1 | y)$  peut être fondée sur les échantillons marginaux  $\{\theta_1^{(s+k)} | y\}$ ,  $k = 1, 2, \dots$ .

Pour les modèles HB intégrés proposés, afin d'obtenir l'estimation a posteriori du taux de chômage, nous mettons en œuvre la méthode d'échantillonnage de Gibbs en produisant des échantillons à partir des lois conditionnelles complètes des paramètres  $\beta$ ,  $\sigma_y^2$  et  $\sigma_{\epsilon}^2$ ,  $u^n$  et  $\theta_i$ . Ces lois conditionnelles complètes sont données en annexe. Les lois gamma inverses qu'il est facile d'échantillonner. Cependant, la loi conditionnelle de  $\theta_i$  n'a pas de forme explicite. Nous utilisons l'algorithme de Metropolis-Hastings dans l'échantillonneur de Gibbs (Chib et Greenberg 1995) pour mettre à jour  $\theta_i$ . À l'instar de You et coll. (2002) et de You et Rao (2002), nous pouvons écrire la loi conditionnelle complète de  $\theta_i$  dans l'échantillonneur de Gibbs sous la forme

$$\theta_i | y, \beta, \sigma_y^2, \sigma_{\epsilon}^2, u^n \propto h(\theta_i) f(\theta_i),$$

le taux de chômage. Nous appliquons l'approche hiérarchique bayésienne au modèle. Nous obtenons les estimations des moyennes et des variances a posteriori en utilisant la méthode d'échantillonnage de Gibbs.

4.1 Modèle hiérarchique bayésien intégré

Nous proposons maintenant le modèle loglinéaire transversal et chronologique intégré sous un cadre hiérarchique bayésien qui suit :

- Conditionnellement à  $\theta_i = (\theta_{i1}, \dots, \theta_{iJ})$ ,  $[y_i | \theta_i] \sim \text{ind}(N(\theta_i), \Sigma_i(\theta_i))$ ;
- Conditionnellement à  $\beta, u_i$  et  $\sigma_i^2$ ,  $[\log(\theta_i) | \beta, u_i, \sigma_i^2] \sim \text{ind}(N(x_i^T \beta + u_i, \sigma_i^2))$ ;
- Conditionnellement à  $u_{i-1}$  et  $\sigma_i^2$ ,  $[u_i | u_{i-1}, \sigma_i^2] \sim \text{ind}(N(u_{i-1}, \sigma_i^2))$ ;
- $\Sigma_i(\theta_i)$  dépend de  $\theta_i$  avec les éléments diagonaux  $\sigma_{ii}^2 = \theta_{ii}^2 (1 - \theta_{ii}) \cdot \tau_i$  et les éléments non diagonaux  $\sigma_{ij}^2 = \rho_{ij}^{1-s} (\bar{\sigma}_{ij}^2 \bar{\sigma}_{ii}^2)$ .
- Marginalement,  $\beta, \sigma_i^2$  et  $\sigma_{ii}^2$  sont mutuellement indépendants, leurs priors étant donnés par  $\beta \propto I$ ,  $\sigma_i^2 \sim \text{IG}(a_1, b_1)$ , et  $\sigma_{ii}^2 \sim \text{IG}(a_2, b_2)$ , où IG dénote une loi gamma inverse et  $a_1, b_1, a_2, b_2$  sont des constantes positives connues auxquelles nous donnons habituellement une valeur très faible pour refléter nos connaissances vagues au sujet de  $\sigma_i^2$  et  $\sigma_{ii}^2$ .

Remarques :

1. Nous utilisons dans le modèle HB proposé un modèle de lien loglinéaire pour le paramètre de petit domaine d'intérêt  $\theta_{ii}^2$ , comme l'ont suggéré You et Rao (2002) et You et Rao (2003).
2. La matrice de covariance d'échantillonnage  $\Sigma_i$  est et coll. (2002) et You et Rao (2002).
3. Nous avons utilisé l'hypothèse d'effets de plan constants pour les petits domaines comme l'ont suggéré Singh, You et Mantel (2005).
4. Le modèle HB propose surmonter les limites du modèle de You et coll. (2000, 2003) en ce qui concerne la modélisation loglinéaire et la spécification de la modélisation de la matrice de covariance d'échantillonnage inconnue. En particulier, nous modélisons cette matrice à l'aide de paramètres de petit domaine  $\theta_{ij}$  en utilisant les estimations lissées des effets de plan pour chaque domaine.

Nous souhaitons estimer le taux de chômage réel  $\theta_{ii}^2$  et, en particulier, le taux de chômage courant  $\theta_{ii}^{IT}$ . Dans l'analyse HB,  $\theta_{ii}^{IT}$  est estimé par sa moyenne a posteriori

respectivement. Alors, la nouvelle matrice de covariance d'échantillonnage lissée  $\bar{\Sigma}_i$  a pour éléments diagonaux  $\bar{\sigma}_{ii}^2$  et pour éléments non diagonaux  $\bar{\sigma}_{ij}^2$ . Cependant, sous cette méthode, la matrice de covariance d'échantillonnage  $\bar{\Sigma}_i$  devient inconnue dans le modèle, puisque  $\sigma_{ii}^2$  et  $\bar{\sigma}_{ii}^2$  dépendent des paramètres inconnus  $\theta_{ii}^2$ , et que ceux-ci sont reliés à un modèle de lien. L'avantage de cette méthode tient au fait que la structure de la matrice de covariance d'échantillonnage est clairement spécifiée dans le modèle. Elle est meilleure que la méthode de lissage en ce sens que la covariance d'échantillonnage est spécifiée explicitement au lieu d'être considérée comme connue.

3.3 Approche de modélisation avec effets de plan égaux

Une autre approche de modélisation est fondée sur l'hypothèse d'effets de plan constants au cours du temps, comme l'ont proposé Singh, You et Mantel (2005), ainsi que Singh, Folsom et Vaish (2005) pour lisser la variance d'échantillonnage  $\sigma_{ii}^2$ . L'effet de plan (deff) pour le domaine au temps  $t$  peut s'écrire approximativement sous la forme

$$\text{deff}^n = \frac{\sigma_{ii}^n}{\theta_{ii}^n (1 - \theta_{ii}^n) / n_{ii}^n}$$

où  $n_{ii}^n$  est la taille de l'échantillon correspondant. Alors, la variance d'échantillonnage  $\sigma_{ii}^2$  peut s'écrire sous la forme  $\sigma_{ii}^2 = \theta_{ii}^n (1 - \theta_{ii}^n) \cdot \text{deff}^n / n_{ii}^n$ . Soit  $\tau_{ii}^n = \text{deff}^n / n_{ii}^n = \sigma_{ii}^n / (\theta_{ii}^n (1 - \theta_{ii}^n))$ . Alors, nous pouvons estimer  $\tau_{ii}^n$  en utilisant les estimations directes de  $\theta_{ii}^n$  et de  $\sigma_{ii}^n$  comme  $\hat{\tau}_{ii}^n = \hat{\sigma}_{ii}^n / (y_{ii}^n (1 - y_{ii}^n))$ . Pour chaque domaine, sous l'hypothèse d'un effet de plan constant et d'une taille d'échantillon constante au cours du temps, nous pouvons obtenir un facteur moyen lissé  $\bar{\tau}_i$  donné par  $\bar{\tau}_i = \sum_{t=1}^T \tau_{ii}^n / T$ . Alors, nous pouvons calculer d'échantillonnage lissée de la forme  $\bar{\sigma}_{ii}^2 = \theta_{ii}^n (1 - \theta_{ii}^n) \cdot \bar{\tau}_i$ , qui de nouveau dépend également de  $\theta_{ii}^n$ . La covariance d'échantillonnage reste de la forme  $\bar{\sigma}_{ij}^2 = \rho_{ij}^{1-s} (\bar{\sigma}_{ij}^2 \bar{\sigma}_{ii}^2)$ , comme dans You et coll. (2003). Soulignons que  $\bar{\tau}_i$  est une moyenne mobile de  $\tau_{ii}^n$  sur la période  $T$  dans le modèle. Toutefois, en pratique, une autre option consiste à utiliser une série chronologique plus longue pour obtenir une estimation plus stable de  $\bar{\tau}_i$  pour chaque domaine, au besoin. Ici, nous utiliserons le modèle à effet de plan constant pour l'estimation du taux de chômage basé sur le facteur à moyenne mobile lissé  $\bar{\tau}_i$ , car nous empruntons de l'information provenant de la période passée  $T$ .

4. Inférence bayésienne hiérarchique

À la présente section, nous proposons un modèle loglinéaire transversal et chronologique intégré pour estimer



You et Rao (2002) ont également étudié le modèle de lien loglinéaire pour le modèle de Fay-Herriot en tant que modèles non appariés d'échantillonnage et de lien avec application à l'estimation du sous-dénominateur au recensement canadien. Les résultats de You et Rao (2002) et de You et coll. (2002) montrent que le modèle de lien loglinéaire donne de fort bons résultats lorsqu'il est appliqué à des problèmes d'estimation sur petits domaines. Ici, nous utiliserons par conséquent le modèle de lien loglinéaire (6) pour estimer le taux réel de chômage  $\theta''$ .

3. Variance d'échantillonnage

En général, nous pouvons obtenir des estimations directes de la variance d'échantillonnage d'après les données d'enquête. Toutefois, ces estimations directes sont instables si les tailles d'échantillon sont faibles. Dans les modèles au niveau du domaine pour l'estimation sur petits domaines, on suppose habituellement que les variances d'échantillonnage sont connues (par exemple, Fay et Herriot 1979; Datta et coll. 1999; You et Rao 2002). Sous cette hypothèse, des estimations fiables (liées) des variances d'échantillonnage sont constituées en utilisant d'autres données et modèles auxiliaires, habituellement en se servant de fonctions de variance généralisées (par exemple, Dick 1995; Datta et coll. 1999). Alternativement, dans le présent article, nous modélisons la matrice de covariance-covariance d'échantillonnage en nous servant des estimations directes de manière particulière qui rend non nécessaire l'hypothèse que les variances et les covariances d'échantillonnage sont connues dans le modèle. Donc, nous simplifions le problème du lissage de la variance d'échantillonnage inconnue et intégrons la modélisation de la variance d'échantillonnage dans le modèle complet.

3.1 Lissage de la matrice de covariance d'échantillonnage

You et coll. (2000, 2003) ont suivi deux étapes pour lisser la matrice de covariance d'échantillonnage. La première consiste à obtenir un coefficient de variation lissé ou constant en calculant pour chaque RMR/CU le coefficient de variation moyen sur une certaine période, que nous dénotons  $CV_i$ , où  $i = 1, 2, \dots, m$ . La deuxième étape consiste à obtenir le coefficient de corrélation avec décalage moyen sur le temps et toutes les RMR/CU, dénoté  $\rho^{(j-s)}$  pour le décalage temporel  $|t-s|$ . Cette étape requiert d'importantes ressources de calcul. Nous avons utilisé des données de l'EPA couvrant une période de trois années (1999 à 2001) pour calculer les coefficients de corrélation (1999 à 2001) pour les valeurs lissées. Dans le modèle, nous traitons les valeurs temporellement et spatialement comme étant les valeurs réelles. Nous obtenons pour le coefficient de corrélation

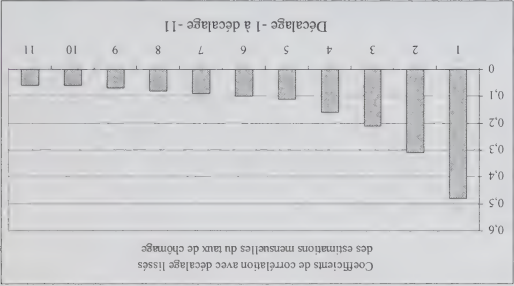


Figure 1 Coefficients de corrélation avec décalage lissés des taux de chômage

3.2 Approche de la modélisation avec CV égaux

Le principal problème de la méthode de You et coll. (2000, 2003) est que les matrices lissées de covariance d'échantillonnage dépendent des estimations directes par sondage  $y''$ , alors que ces dernières ne sont pas fiables pour certaines petites régions. Il convient de souligner que la variance d'échantillonnage réelle peut s'écrire sous la forme  $\sigma_v^2 = \theta_v^2 (CV'')^2$ . Partant de l'hypothèse d'un CV constant au cours du temps pour un domaine donné, You et coll. (2003) ont proposé dans leurs conclusions d'utiliser des estimations de la forme  $\hat{\sigma}_v^2 = \theta_v^2 (CV')^2$ . Par conséquent, les covariances et les covariances lissées,  $\hat{\rho}^{(j-s)}(\hat{\sigma}_v'')$  pour les variances et les covariances lissées,

d'échantillonnage. Le modèle de lien pour le taux de chômage réel  $\theta''$  peut s'écrire sous la forme

$$\theta'' = x''\beta + v_i, \quad i = 1, \dots, m, \quad (2)$$

où  $x''$  est la variable auxiliaire et  $v_i$  est l'effet aléatoire particulier au domaine. Pour chaque point dans le temps (chaque mois), nous pouvons utiliser le modèle de Fay-Herriot pour produire les estimations directes mensuelles. Ce modèle combine l'information transversale, mais n'est pas renforcé par emprunt d'information recueillie au cours des périodes antérieures.

## 2.2 Modèle transversal et chronologique

Étant donné le plan d'échantillonnage du processus de renouvellement de l'échantillon de l'EPA, il existe, dans

chaque domaine, un chevauchement important d'échantillons sur des périodes de six mois. Par conséquent, pour un domaine donné  $i$ , il est nécessaire de prendre en compte la corrélation entre les erreurs d'échantillonnage  $e''$  et  $e''_s$  ( $i \neq s$ ). You et coll. (2000, 2003) ont proposé un modèle transversal et chronologique pour estimer les taux de chômage d'après l'EPA. Ils n'ont utilisé que les données recueillies au cours des six mois précédents pour prédire le taux du mois courant, puisque le renouvellement de l'échantillon de l'EPA est fondé sur un cycle de six mois. Chaque mois, le sixième de l'échantillon de l'EPA est remplacé. Donc, après six mois, la corrélation entre les estimations est faible (voir la section 2.1 pour les coefficients de corrélation retardés). Soit  $y_i = (y_i^{(1)}, \dots, y_i^{(T)})'$ ,  $\theta_i = (\theta_i^{(1)}, \dots, \theta_i^{(T)})'$ , et  $e_i = (e_i^{(1)}, \dots, e_i^{(T)})'$ , où  $T = 6$  ici. En supposant que  $e_i$  suit une loi normale multivariée de vecteur moyenne 0 et de matrice de covariance d'échantillonnage  $\Sigma_i$ , nous avons

$$y_i \sim N(\theta_i, \Sigma_i), \quad i = 1, \dots, m.$$

Donc nous supposons que  $y_i$  est sans biais par rapport au plan pour  $\theta_i$ . La matrice de covariance d'échantillonnage  $\Sigma_i$  est inconnue dans le modèle. Des estimations directes des matrices de covariance d'échantillonnage sont disponibles. Dans le contexte de l'estimation sur petits domaines fondée sur un modèle au niveau du domaine, on suppose habituellement que la variance d'échantillonnage est connue dans le modèle. Habituellement, on utilise un estimateur lissé de la variance d'échantillonnage pour aborder le problème; par exemple, une autre approche pour aborder le problème, ainsi que voir Wang et Fuller (2003), You et Dick (2004), ainsi que You et Chapman (2006). En ce qui concerne l'estimation des taux de chômage, nous fournissons des précisions sur le

où  $v_i$  est un effet aléatoire de domaine que nous supposons suivre une loi  $N(0, \sigma_v^2)$  et  $u''_i$  est une composante temporelle (période) et spatiale (domaine) aléatoire. Nous pouvons en outre supposer que  $u''$  suit un processus de marche aléatoire sur les périodes  $t = 1, \dots, T$ , autrement dit que

$$u''_t = u''_{t-1} + \varepsilon''_t \quad (4)$$

où  $\varepsilon'' \sim N(0, \sigma_\varepsilon^2)$ . Alors,  $\text{cov}(u''_t, u''_s) = \min(t, s) \sigma_\varepsilon^2$ . Le vecteur de régression  $\beta$  et les composantes de la variance  $\sigma_v^2$  et  $\sigma_\varepsilon^2$  sont inconnus dans le modèle et doivent être estimés. En combinant les modèles (1), (3) et (4), nous obtenons un modèle mixte linéaire avec composantes temporelles de la forme

$$y'' = x''\beta + v_i + u'' + e''_i, \quad i = 1, \dots, m, \quad t = 1, \dots, T. \quad (5)$$

You et coll. (2003) ont montré que le modèle transversal et chronologique (5) est supérieur au modèle de Fay-Herriot en ce qui concerne le lissage des estimations directes et la réduction des coefficients de variation des estimations directes pour l'estimation du taux de chômage d'après l'EPA.

Nous avons utilisé un modèle à marche aléatoire pour stationnaire pour  $u''$ . You et coll. (2003) ont montré que le modèle à marche aléatoire sur  $u''$  fournissait un modèle mieux ajusté à l'estimation du taux de chômage que le modèle autorégressif AR(1). Data et coll. (1999) ont également utilisé un modèle à marche aléatoire pour estimer les taux de chômage au niveau de l'État aux États-Unis.

## 2.3 Modèle de lien loglinéaire

Cependant, l'une des limites du modèle (3) est due au fait que le modèle de lien pour le paramètre d'intérêt, c'est-à-dire le taux réel de chômage  $\theta''$ , est un modèle linéaire avec effets aléatoires normaux. Puisque  $\theta''$  est un nombre positif compris entre 0 et 1, et qu'il est proche de 0, le modèle de lien linéaire avec effets aléatoires normaux pourrait introduire des estimations négatives de  $\theta''$  pour certains petits domaines. Afin d'éviter ce problème, You, Chen et Gambino (2002) ont proposé un modèle de lien loglinéaire pour  $\theta''$  de la forme suivante :

$$\log(\theta'') = x''\beta + v_i + u''_i, \quad i = 1, \dots, m, \quad t = 1, \dots, T. \quad (6)$$



chômage d'après l'EPA sous un cadre hiérarchique bayésien (HB). L'idée est de modéliser ensemble les paramètres d'intérêt et les variances d'échantillonnage, comme l'ont suggéré You et coll. (2003), ainsi que You et Chapman (2006). Nous appliquons le modèle proposé aux données de l'EPA de 2005 et obtenons les estimations des taux de chômage fondés sur un modèle. Nous donnons aussi une comparaison des estimations HB aux estimations directes d'après l'EPA et l'analyse de l'ajustement du modèle.

Le plan de l'article est le suivant. À la section 2, nous présentons divers modèles pour petits domaines proposés dans la littérature pour l'estimation des taux de chômage et nous en discutons. À la section 3, nous examinons le problème du lissage et de la modélisation de la matrice de covariance d'échantillonnage. À la section 4, nous proposons un modèle non linéaire intégré à effets mixtes dans un cadre hiérarchique bayésien et décrivons l'utilisation de l'échantillonnage de Gibbs pour générer des échantillons à partir de la loi conjointe à posteriori. À la section 5, nous appliquons le modèle proposé aux données de l'EPA et obtenons les estimations HB pour les taux de chômage pour petits domaines. Nous donnons aussi l'analyse et l'évaluation du modèle. Enfin, à la section 6, nous présentons certaines conclusions et discussions de l'orientation de futurs travaux.

## 2. Modèles pour petits domaines

### 2.1 Modèle transversal

Les modèles transversaux, ou au niveau du domaine, sont utilisés pour produire des estimations fondées sur un modèle fiable par combinaison de l'information auxiliaire au niveau du domaine et des estimations directes au niveau du domaine. Un modèle transversal de base est celui, bien connu, de Fay-Herriot (Fay et Herriot 1979). Il possède deux composantes, à savoir 1) un modèle d'échantillonnage pour les estimations directes par sondage et 2) un modèle de lien qui relie les paramètres du petit domaine à des variables auxiliaires au niveau du domaine grâce à un modèle de régression linéaire. Pour l'estimation mensuelle du taux de chômage d'après l'EPA, dénotons par  $\theta_i$  le taux de chômage réel pour la  $i^{\text{e}}$  RMR/CU à un point particulier dans le temps (mois)  $t$ , où  $i = 1, \dots, m$ , où  $m$  est le nombre de RMR/CU, et soit  $y_i^t$  l'estimation directe d'après l'EPA de  $\theta_i$ . Alors, le modèle d'échantillonnage pour  $y_i^t$  peut être exprimé sous la forme

$$y_i^t = \theta_i + e_i^t, \quad i = 1, \dots, m, \quad (1)$$

où  $e_i^t$  est l'erreur d'échantillonnage associée à l'estimateur direct  $y_i^t$ . Nous supposons que cette erreur suit une loi normale telle que  $e_i^t \sim N(0, \sigma_e^2)$  où  $\sigma_e^2$  est la variance

En général, on utilise les estimateurs directs, fondés uniquement sur des données d'échantillon sur un domaine particulier, pour estimer les paramètres pour de grands domaines, tels que le pays et les provinces. Par contre, les tailles d'échantillon pour les petits domaines, particulièrement les petites régions géographiques, sont rarement suffisantes pour fournir des estimations directes fiables. Quand on calcule des estimations pour des petits domaines, il est nécessaire d'emprunter de l'information supplémentaire à des domaines apparentés pour former des estimateurs indirects qui accroissent la taille effective d'échantillon et, donc, la précision. De tels estimateurs indirects sont fondés sur des modèles implicites ou explicites qui établissent un lien avec des petits domaines apparentés grâce à des données supplémentaires, tels que des dénombrements du recensement et des dossiers administratifs. Aujourd'hui il est généralement reconnu que, s'il faut recourir à des estimations indirectes, celles-ci devraient être fondées sur des modèles explicites qui relient les petits domaines d'intérêt grâce à des données supplémentaires; voir Rao (2003). Les estimateurs fondés sur un modèle sont des estimations indirectes en ce sens qu'ils sont obtenus en utilisant des estimations pour petits domaines, des estimations directes et des variables auxiliaires connexes. Les estimateurs fondés sur un modèle sont établis en vue d'améliorer les estimateurs directs fondés sur le plan de sondage en ce qui concerne la précision et la fiabilité, c'est-à-dire la réduction des coefficients de variation. Habituellement, les estimateurs sur petits domaines sont renforcés par emprunt de données recueillies pour des petits domaines similaires à une période donnée ou pour le même domaine au fil du temps, mais non les deux. Ces dernières années, plusieurs méthodes ont été élaborées en vue d'emprunter simultanément des données de renfort transversales (dimension spatiale) et longitudinales (dimension temporelle). Les estimateurs fondés sur l'approche établie par Rao et Yu (1994), Ghosh, Nangia et Kim (1996), Datta, Lahiri, Maity et Lu (1999), Datta, Lahiri et Maity (2002), ainsi que You, Rao et Gambino (2000, 2003), permettent d'exploiter simultanément les deux dimensions pour produire des estimations améliorées ayant des propriétés souhaitables pour les petits domaines. En particulier, You et coll. (2000, 2003) ont étudié l'estimation fondée sur un modèle de chômage pour des régions infra-provinciales locales, comme les RMR et les agglomérations de recensement (AR) au Canada. Ils ont obtenu des estimateurs fondés sur un modèle efficaces et un ajustement adéquat des modèles pour l'estimation des taux de chômage d'après l'EPA. Toutefois, le modèle proposé par You et coll. (2000, 2003) présente certaines limites. Dans le présent article, nous discutons de ces limites et proposons un nouveau modèle intégré pour l'estimation des taux de



# Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions infraprovinciales au Canada

Yong You<sup>1</sup>

## Résumé

L'Enquête sur la population active (EPA) réalisée au Canada permet de produire des estimations mensuelles du taux de chômage aux niveaux national et provincial. Le programme de l'EPA diffuse aussi des estimations du chômage pour des régions infraprovinciales, comme les régions métropolitaines de recensement (RMR) et les centres urbains (CU). Cependant, pour certaines de ces régions infraprovinciales, les estimations directes ne sont pas fiables, parce que la taille de l'échantillon est assez petite. Dans le contexte de l'EPA, l'estimation pour de petites régions a trait à l'estimation des taux de chômage pour des régions infraprovinciales telles que les RMR/CU à l'aide de modèles pour petits domaines. Dans le présent article, nous discutons de divers modèles, dont celui de Fay-Herriot et des modèles transversaux ainsi que bayésien (HB) pour l'estimation du taux de chômage d'après les données de l'EPA. Nous utilisons les données mensuelles sur les bénéficiaires de l'assurance-emploi (a-e.) au niveau de la RMR ou du CU comme covariables auxiliaires dans le modèle. Nous appliquons une approche HB ainsi que la méthode d'échantillonnage de Gibbs pour obtenir les estimations des moyennes et des variances a posteriori des taux de chômage au niveau de la RMR ou du CU. Le modèle HB propose produit des estimations fondées sur un modèle fiables si l'on s'en tient à la réduction du coefficient de variation. Nous présentons dans l'article une analyse d'ajustement du modèle et une comparaison des estimations fondées sur le modèle aux estimations directes.

Mots clés : Effet de plan; cadre hiérarchique bayésien; modèle logarithmique à effets mixtes; vérification du modèle; variance d'échantillonnage; petit domaine.

## 1. Introduction

Le taux de chômage est généralement considéré comme un indicateur important des résultats économiques. Au Canada, les estimations du taux de chômage sont produites mensuellement par le programme de l'Enquête sur la population active (EPA) de Statistique Canada. Cette dernière est une enquête mensuelle réalisée auprès de 53 000 ménages sélectionnés conformément à un plan d'échantillonnage stratifié à plusieurs degrés. Chaque mois, un sixième de l'échantillon est remplacé. Donc, cinq sixièmes de l'échantillon sont communs à deux mois consécutifs. Ce chevauchement des échantillons crée des corrélations qui peuvent être exploitées pour produire de meilleures estimations au moyen d'autres méthodes que l'estimation directe, telles que les méthodes fondées sur un modèle, qui renforcent l'estimation à l'aide de données accumulées au fil du temps; cette approche sera examinée plus en détail à la section 2. Pour une description détaillée du plan de sondage de l'EPA, consulter Gambino, Singh, Dufour, Kennedy et Lindeyer (1998). Le programme de l'EPA diffuse des estimations mensuelles du taux de chômage pour de grandes régions, telles que l'ensemble du pays et les provinces, ainsi que pour des régions locales (petites régions ou petits domaines), telles que les régions métropolitaines de recensement (RMR), c'est-à-dire les villes comptant plus de 100 000 habitants) et les autres centres

urbains du Canada. Bien que les estimations nationales et provinciales soient celles qui intéressent le plus les médias, les estimations infraprovinciales sont également très importantes. Le programme de l'assurance-emploi (a-e.) les utilise pour établir les règles à appliquer pour administrer le programme. En outre, les taux de chômage dans les RMR et les CU sont examinés de près au niveau local. Cependant, pour nombre de régions locales, la taille de l'échantillon n'est pas suffisante pour calculer des estimations directes adéquates, car l'EPA est conçue pour produire des estimations adéquates ou fiables aux niveaux national et provincial. Pour les estimations nationales, le coefficient de variation (CV) est de l'ordre de 2 % et pour les estimations provinciales, de l'ordre de 4 % à 7 %. En revanche, pour les RMR et les CU, il varie d'environ 7 % à 50 %. Les pour certains CU, il est même supérieur à 50 %. Les estimations directes basées sur les données de l'EPA ne sont pas très fiables pour certaines régions municipales, le CV étant très grand à cause de la petite taille de l'échantillon pour ces régions. Par conséquent, d'autres estimateurs, en particulier des estimateurs fondés sur un modèle, sont envisagés pour améliorer les estimations directes d'après l'EPA pour les petites régions. L'objectif du présent article est d'obtenir un estimateur fondé sur un modèle fiable qui produit de meilleures estimations que l'estimateur direct d'après l'EPA, c'est-à-dire des estimations dont le CV est petit et stable.

Il découle maintenant de (A.1), (A.3) et (A.16) que

$$E\hat{Q}M(\hat{\mu}_{LV}^i) = E\hat{Q}M(\hat{\mu}_B^i)$$

$$+ \left\{ \hat{x}_i' \Sigma_v X_i' R_i^{-1} X_i \Sigma_v \hat{x}_i - \hat{x}_i' \Sigma_v \hat{x}_i + \frac{\sigma^2}{n_i} \right\},$$

Lehtonen, R., Särndal, C.-E. et Veijanen, A. (2003), L'effet du choix d'un modèle dans l'estimation par domaine, dont les petits domaines, *Techniques d'enquête*, 29, 37-49.

Mourá, F.A.S., et Holt, D. (1999), Production d'estimations régionales à partir de modèles multiniveaux, *Techniques d'enquête*, 25, 81-89.

comme il est énoncé au théorème 3.

Bibliographie

Rao, J.N.K. (2003). *Small Area Estimation*. Hoboken : New York : John Wiley & Sons, Inc.

Särndal, C.-E., Swensson, B. et Wretman, J.H. (1992). *Model Assisted Survey Sampling*. New York : Springer-Verlag.

Searle, S.R. (1971). *Linear Models*. New York : John Wiley & Sons, Inc.

Lehtonen, R., et Veijanen, A. (1999). Domain estimation with logistic generalized regression and related estimators. *IASS Satellite Conference on Small Area Estimation*, Riga : Latvian Council of Science, 121-128.

LEMME 1 (Searle 1971). Si  $y$  est un vecteur de dimension  $n \times 1$  de moyenne  $\mu$  et de matrice de variance-covariance  $\Sigma$  et que  $b$  est un vecteur de dimension  $n \times 1$ , alors  $E(yby') = \Sigma b + \mu\mu'$ .

Donc,

$$E(\underline{v}_i' - \underline{x}_i' Z_i \beta)(\underline{x}_i' \underline{v}_i) = \underline{x}_i' \Sigma_v \underline{x}_i - \underline{x}_i' Z_i \beta \underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta$$

Nous pouvons alors écrire (A.2) sous la forme :

$$E[\underline{v}_i' - \underline{x}_i' (Z_i \beta + \underline{v}_i)]^2 = \underline{x}_i' \Sigma_v \underline{x}_i + \frac{\sigma_v^2}{2}$$

$$+ \underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \Sigma_v \underline{x}_i - 2 \underline{x}_i' \Sigma_v \underline{x}_i = \underline{x}_i' \Sigma_v X_i' V_i^{-1} \underline{x}_i - \underline{x}_i' \Sigma_v \underline{x}_i + \frac{\sigma_v^2}{2} \quad (\text{A.3})$$

Enfin, nous devons trouver le terme produit croisé de

(A.1). Nous avons

$$E[(\underline{v}_i' - \underline{x}_i' (Z_i \beta + \underline{v}_i))(\underline{x}_i' (\underline{v}_i - \underline{v}_i))] = E[\underline{v}_i' X_i' (\underline{v}_i' - \underline{v}_i')] =$$

$$- E[\underline{x}_i' (Z_i \beta + \underline{v}_i) X_i' (\underline{v}_i' - \underline{v}_i')]. \quad (\text{A.4})$$

où le premier terme du deuxième membre de (A.4) peut s'écrire

$$E[\underline{v}_i' X_i' (\underline{v}_i' - \underline{v}_i')] = E[\underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} (\underline{v}_i' - \underline{v}_i')] = E[\underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} \underline{v}_i']$$

$$- E[\underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta]$$

$$- E[\underline{v}_i' X_i' \underline{v}_i'] \quad (\text{A.5})$$

Le premier terme de (A.5) est

$$E[\underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} \underline{v}_i'] = E\left[\frac{1}{L} \underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} \underline{v}_i'\right] = \frac{1}{L} X_i' (\Sigma_v X_i' + Z_i' \beta X_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta) \quad (\text{A.6})$$

le deuxième terme de (A.5) peut s'écrire

$$E[\underline{v}_i' X_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta] = \underline{x}_i' Z_i \beta X_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta, \quad (\text{A.7})$$

et le troisième terme peut être exprimé par

$$E[(\underline{v}_i' - \underline{x}_i' (Z_i \beta + \underline{v}_i)) X_i' (\underline{v}_i' - \underline{v}_i')] = 0. \quad (\text{A.16})$$

Donc, il découle de (A.4), (A.9) et (A.15) que

$$E[\underline{x}_i' (Z_i \beta + \underline{v}_i) X_i' (\underline{v}_i' - \underline{v}_i')] = 0. \quad (\text{A.15})$$

substitution dans (A.10), nous obtenons

Par conséquent, en introduisant (A.11) à (A.14) par

$$= \underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \Sigma_v \underline{x}_i. \quad (\text{A.14})$$

$$= E[\underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i \underline{v}_i' \underline{v}_i']$$

$$= E[\underline{x}_i' \Sigma_v X_i' V_i^{-1} (X_i Z_i \beta + X_i \underline{v}_i' + e_i) X_i' \underline{v}_i']$$

$$- E[\underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i Z_i \beta X_i' \underline{v}_i']$$

$$= E[\underline{x}_i' \Sigma_v X_i' V_i^{-1} \underline{v}_i' X_i' \underline{v}_i']$$

$$= E[\underline{x}_i' \Sigma_v X_i' V_i^{-1} (\underline{v}_i' - X_i' Z_i \beta) X_i' \underline{v}_i']$$

$$E[\underline{x}_i' \underline{v}_i' X_i' \underline{v}_i']$$

et

$$E[\underline{x}_i' Z_i \beta X_i' \underline{v}_i'] = 0 \quad (\text{A.13})$$

$$= \underline{x}_i' \Sigma_v X_i' V_i^{-1} X_i \underline{v}_i' \Sigma_v \underline{x}_i \quad (\text{A.12})$$

$$\times \underline{x}_i' E(\Sigma_v X_i' V_i^{-1} (\underline{v}_i' - X_i' Z_i \beta))$$

$$+ \underline{x}_i' E(\Sigma_v X_i' V_i^{-1} (\underline{v}_i' - X_i' Z_i \beta))$$

$$= \underline{x}_i' \text{Var}[\Sigma_v X_i' V_i^{-1} (\underline{v}_i' - X_i' Z_i \beta)] X_i'$$

$$E[\underline{x}_i' \underline{v}_i' X_i' \underline{v}_i'] = \underline{x}_i' \Sigma_v \underline{x}_i + \underline{x}_i' E(\underline{v}_i') X_i' E(\underline{v}_i')$$

$$= E[\underline{x}_i' Z_i \beta X_i' \Sigma_v X_i' V_i^{-1} (\underline{v}_i' - X_i' Z_i \beta)] = 0, \quad (\text{A.11})$$

$$E[\underline{x}_i' Z_i \beta X_i' \underline{v}_i']$$

quatre termes du deuxième membre de (A.10) :

Alors, nous obtenons l'expression qui suit pour les

$$- E[\underline{x}_i' Z_i \beta X_i' \underline{v}_i'] - E[\underline{x}_i' \underline{v}_i' X_i' \underline{v}_i'] \quad (\text{A.10})$$

$$= E[\underline{x}_i' Z_i \beta X_i' \underline{v}_i'] + E[\underline{x}_i' \underline{v}_i' X_i' \underline{v}_i']$$

$$E[\underline{x}_i' (Z_i \beta + \underline{v}_i) X_i' (\underline{v}_i' - \underline{v}_i')]$$

de (A.4). Nous avons

Nous nous penchons maintenant sur le deuxième terme

$$E[\underline{v}_i' X_i' (\underline{v}_i' - \underline{v}_i')] = 0. \quad (\text{A.9})$$

(A.5), nous avons

Par conséquent, en substituant (A.6), (A.7) et (A.8) dans

$$= E[\underline{x}_i' \underline{v}_i' X_i' \underline{v}_i'] = \underline{x}_i' \Sigma_v \underline{x}_i \quad (\text{A.8})$$

$$E[\underline{v}_i' X_i' \underline{v}_i'] = E[(\underline{x}_i' Z_i \beta + \underline{x}_i' \underline{v}_i' + e_i) X_i' \underline{v}_i']$$



Tableau 6  
Comparaison des estimateurs sur petits domaines : EQM  
moyenne ( $EQM_1$ ) et erreur relative absolue moyenne (ERA2)  
sous un cadre de population finie ( $n_f = 10$ )

Mesure	Modèle sans z	Modèle avec z
de la	ERG Nouveau	ERG Nouveau
qualité	ERG	ERG
ERA2	0,20	0,18
EQM2	6,53	6,77
	4,73	6,75
	5,24	5,24

5. Sommaire

Dans le présent article, nous avons calculé l'erreur quadratique moyenne (EQM) de modélisation pour un nouveau estimateur GREG assisté par un modèle à deux niveaux d'une moyenne de petit domaine proposé par Lehtonen et Veijanen (1999). En outre, nous avons utilisé un ensemble de données de Moura et Holt (1999) pour démontrer empiriquement que l'estimateur BLUP est considérablement plus efficace que le nouvel estimateur GREG en ce qui concerne l'EQM sous le modèle, tandis que le nouvel estimateur GREG n'est que légèrement plus efficace que l'estimateur GREG habituel fondé sur le modèle de régression  $y_i = X_i'Z_i\beta + e_i$ ,  $i = 1, \dots, m$ . En outre, à l'aide d'une étude par simulation sous un cadre fondé sur un modèle, nous avons montré que le nouvel estimateur GREG a systématiquement de meilleures propriétés que l'estimateur GREG habituel en ce qui concerne l'EQM, et l'erreur relative absolue moyenne, ERAM. Cependant, étant donné l'emprunt d'information à des petits domaines apparents, l'estimateur EBLUP donne de nettement meilleurs résultats que les estimateurs GREG habituel et nouveau. En outre, nous avons réalisé une étude par simulation sous un cadre de population finie et montré que l'estimateur EBLUP surpasse le nouvel estimateur GREG et l'estimateur GREG habituel en ce qui concerne l'EQM et l'ERAM.

Remerciements

La présente étude a été financée par une subvention du Conseil de recherches en sciences naturelles et en génie du Canada. Elle est fondée sur un chapitre de la thèse de doctorat de M.Torabi rédigée sous la supervision de J.N.K. Rao. Les auteurs remercient les examinateurs et le rédacteur adjoint de leurs commentaires constructifs concernant la version originale du présent article.

Annexe

Calcul de  $EQM(\mu_{LV}^i)$  :

$$\begin{aligned} EQM(\mu_{LV}^i) &= E[(\mu_{LV}^i - \mu_i)^2] \\ &= E[X_i'(\hat{v}_i - v_i)]^2 + E[\hat{v}_i - v_i | Z_i\beta + \hat{v}_i]^2 \\ &\quad + 2E[(\hat{v}_i - v_i)(Z_i\beta + \hat{v}_i)]X_i'(\hat{v}_i - v_i), \quad (A.1) \\ &\quad + E(\hat{v}_i^2) - 2E[\hat{v}_i(Z_i\beta + \hat{v}_i)]X_i'(\hat{v}_i - v_i), \quad (A.2) \\ E[\hat{v}_i - v_i | Z_i\beta + \hat{v}_i]^2 &= E[\hat{v}_i - v_i]^2 \\ &\quad + E(\hat{v}_i^2) - 2E[\hat{v}_i(Z_i\beta + \hat{v}_i)]X_i'(\hat{v}_i - v_i), \quad (A.2) \end{aligned}$$

où

$$E[\hat{v}_i - v_i | Z_i\beta + \hat{v}_i]^2 = \text{Var}(\hat{v}_i) = \hat{v}_i' \Sigma_v \hat{v}_i + \frac{\sigma_e^2}{2}$$

et

$$E(\hat{v}_i^2) = \text{Var}(\hat{v}_i) + [E(\hat{v}_i^2)]^2$$
$$= \text{Var}[\hat{v}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i' Z_i \beta)]$$
$$+ [E(\hat{v}_i' \Sigma_v X_i' V_i^{-1} (y_i - X_i' Z_i \beta))]^2$$
$$= \hat{v}_i' \Sigma_v X_i' V_i^{-1} X_i' \Sigma_v \hat{v}_i$$

En outre,

$$E[(\hat{v}_i - v_i)(Z_i\beta + \hat{v}_i)] = E[\hat{v}_i' \Sigma_v X_i' V_i^{-1} y_i]$$

où le deuxième terme est nul. Par conséquent, nous pouvons écrire

$$E[(\hat{v}_i - v_i)(Z_i\beta + \hat{v}_i)] = E[\hat{v}_i' \Sigma_v X_i' V_i^{-1} y_i]$$

où le premier terme peut s'écrire

$$E[\hat{v}_i' \Sigma_v X_i' V_i^{-1} y_i] = E[\hat{v}_i' \Sigma_v X_i' V_i^{-1} X_i' Z_i \beta + \hat{v}_i' \Sigma_v X_i' V_i^{-1} X_i' Z_i \beta]$$

en utilisant le lemme suivant :

où  $\hat{\mu}_i^{(b)}$  dénote  $\hat{\mu}_i^{(b)}$ ,  $\hat{\mu}_i^{(b)}$  ou  $\hat{\mu}_i^{(b)}$ . Nous présentons les résultats au tableau 4. Les deux modèles avec covariable  $z$  au niveau du domaine et sans covariable  $z$  produisent des valeurs un peu plus faibles de  $\hat{EQM}_1$  et de  $\overline{ERA}_1$  pour le nouvel estimateur GREG que pour l'estimateur GREG habituel. Cependant,  $\hat{EQM}_1$  et  $\overline{ERA}_1$  sont significativement plus faibles pour l'estimateur EBLUP, à cause de l'emprunt d'information aux domaines apparentés. En outre, la comparaison des tableaux 2 et 4 révèle que les valeurs de  $\hat{EQM}_1$  dans le tableau 4 sont un peu plus grandes que les valeurs correspondantes dans le tableau 2, à cause de l'estimation des paramètres du modèle.

Tableau 4  
Comparaison des estimateurs sur petits domaines :  $\hat{EQM}$  moyenne ( $\hat{EQM}_1$ ) et erreur relative absolue moyenne ( $\overline{ERA}_1$ ) sous un cadre fondé sur un modèle

Mesure	Modèle sans z	GREG	GREG	EBLUP	GREG	EBLUP
qualité de la						
$\hat{EQM}_1$	1,93	1,73	0,67	1,84	1,73	0,73
$\overline{ERA}_1$	0,14	0,13	0,08	0,13	0,12	0,08

## 4.2.2 Étude par simulation sous un cadre de population finie

Pour étudier les propriétés des estimateurs sous un cadre de population finie, nous avons créé une population finie synthétique d'après les données brésiliennes constituées de  $n = 3\,876$  valeurs d'échantillon  $(y_{ij}^{(b)}, x_{1ij}^{(b)}, x_{2ij}^{(b)}, z_i)$ . En reproduisant les valeurs d'échantillon  $(y_{ij}^{(b)}, x_{1ij}^{(b)}, x_{2ij}^{(b)}, z_i)$  cinq fois, nous avons obtenu le nouvel ensemble de données  $(y, x_1, x_2, z)$  de taille 19 380 que nous avons traité comme notre population réelle.

Nous avons produit 500 échantillons indépendants ( $B = 500$ ), chacun de taille  $n = 700$  et  $n = 1\,400$ , en tirant des échantillons aléatoires simples de taille  $n_1 = 5$  et  $n_2 = 10$ , dans chaque domaine  $i = 1, \dots, 140$ . Comme auparavant, pour chaque échantillon, nous avons calculé  $\hat{\beta}^{(b)}$  pour la nouvelle méthode GREG et pour la méthode EBLUP, et  $\hat{\beta}^{(b)}$  pour la méthode GREG. En outre, nous avons calculé  $\hat{\Sigma}^{(b)}$  et  $\hat{\sigma}_{2(b)}^2$  par la méthode du REML. Nous avons également calculé la moyenne de population de  $y_{ij}$  pour chaque domaine  $i$  par

$$\bar{y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i, \quad i = 1, \dots, 140,$$

où  $N_i$  est la taille de la population dans le  $i^{\text{e}}$  domaine. En outre, pour chaque échantillon  $b = 1, \dots, B$ , nous avons calculé la nouvelle estimation GREG de la moyenne du  $i^{\text{e}}$  domaine selon

$$\hat{y}_i^{(b)} = \hat{y}_i^{(b)} + (X_i' - X_i^{(b)'}) (Z_i' \hat{\beta}^{(b)} + \hat{v}_i^{(b)}),$$

L'estimation GREG selon  $\hat{y}_i^{(b)} = \hat{y}_i^{(b)} + (X_i' - X_i^{(b)'}) (Z_i' \hat{\beta}^{(b)} + \hat{v}_i^{(b)})$  et l'estimation EBLUP selon  $\hat{y}_i^{(b)} = \hat{y}_i^{(b)} + (X_i' - X_i^{(b)'}) (Z_i' \hat{\beta}^{(b)} + \hat{v}_i^{(b)})$ , où

$$f_i = n_i / N_i, \quad X_i^* = \frac{N_i \bar{X}_i - n_i \bar{X}_i}{N_i - n_i},$$

Nous avons calculé la moyenne des erreurs quadratiques moyennes ( $\hat{EQM}_2$ ) et la moyenne des erreurs relatives absolues ( $\overline{ERA}_2$ ) comme étant

$$\hat{EQM}_2 = \frac{1}{m} \sum_{i=1}^m \hat{EQM}_{2i}, \quad \overline{ERA}_2 = \frac{1}{m} \sum_{i=1}^m \overline{ERA}_{2i},$$

où

$$\hat{EQM}_{2i} = \frac{1}{B} \sum_{b=1}^B (\hat{y}_i^{(b)} - \bar{y}_i)^2, \quad \overline{ERA}_{2i} = \frac{1}{B} \sum_{b=1}^B |\hat{y}_i^{(b)} - \bar{y}_i| / \bar{y}_i,$$

et  $\hat{y}_i^{(b)}$  dénote  $\hat{y}_i^{(b)}$ ,  $\hat{y}_i^{(b)}$  ou  $\hat{y}_i^{(b)}$ . Nous présentons les résultats aux tableaux 5 et 6 pour  $n_1 = 5$  et  $n_2 = 10$ , respectivement. Nous envisageons les modèles avec covariable au niveau du domaine  $z$  et sans la covariable  $z$ .

Tableau 5  
Comparaison des estimateurs sur petits domaines :  $\hat{EQM}$  moyenne ( $\hat{EQM}_2$ ) et erreur relative absolue moyenne ( $\overline{ERA}_2$ ) sous un cadre de population finie ( $n_1 = 5$ )

Mesure	Modèles sans z	GREG	GREG	EBLUP	GREG	EBLUP
qualité de la						
$\hat{EQM}_2$	11,03	10,02	6,50	10,76	10,06	7,06
$\overline{ERA}_2$	0,27	0,24	0,18	0,25	0,23	0,22

Le tableau 5 montre que, pour  $n_1 = 5$ , le nouvel estimateur GREG donne de meilleurs résultats que l'estimateur GREG habituel en ce sens que  $\hat{EQM}_2$  et  $\overline{ERA}_2$  sont plus faibles. Par ailleurs, le tableau 6 révèle que, pour  $n_2 = 10$ , l'estimateur GREG habituel a d'un peu meilleures propriétés que le nouvel estimateur GREG en ce qui concerne  $\hat{EQM}_2$ , mais non  $\overline{ERA}_2$ . Cependant, l'estimateur EBLUP donne des valeurs considérablement plus faibles de  $\hat{EQM}_2$  et  $\overline{ERA}_2$  que l'estimateur GREG habituel et le nouvel estimateur GREG dans les deux cas, parce qu'il emprunte de l'information aux petits domaines apparentés. Par exemple, pour le modèle sans  $z$  et  $n_1 = 5$ ,  $\hat{EQM}_2 = 10,02$ , 11,03 et 6,50 pour le nouvel estimateur GREG, l'estimateur GREG habituel et l'estimateur EBLUP, respectivement.

**Tableau 3**  
Comparaison des estimateurs sur petits domaines : efficacité relative (EFF) et EQM moyenne (EQM) pour le cas d'une matrice de covariance générale basé sur l'ensemble de données de Moura et Holt (1999)

Mesure de la qualité	Modèle sans z	Nouveau GREG	BLUP
EFF	100 %	108 %	253 %
EQM	2.02	1.87	0.80

**4.2 Étude par simulation**

**4.2.1 Étude par simulation sous un cadre fondé sur un modèle**

Afin d'étudier l'efficacité du nouvel estimateur GREG lorsque les paramètres du modèle sont estimés, nous avons entrepris une petite étude par simulation fondée sur les modèles à deux niveaux (11)-(12) et (13)-(14). Nous avons considéré uniquement la structure de covariance diagonale  $\Sigma_v$  avec les éléments diagonaux  $\sigma_1^2$  et  $\sigma_2^2$ . Nous avons de nouveau utilisé les données tirées de Moura et Holt (1999). Les estimations de  $\beta_0, \beta_1, \beta_2, \alpha_1, \alpha_2, \sigma_1^2, \sigma_2^2$  et  $\sigma_e^2$  présentées au tableau 1 sont traitées comme les valeurs réelles.

Dans notre étude par simulation, nous avons pris  $(x_{1ij}, x_{2ij}, z_i)$  provenant de Moura et Holt (1999), puis nous avons généré  $y_{ij}$  en nous fondant sur les modèles (11)-(12) et (13)-(14). En utilisant les échantillons ainsi produits  $(y_{ij}^{(b)}, x_{1ij}^{(b)}, x_{2ij}^{(b)}, z_i^{(b)})$ ,  $b = 1, \dots, B = 1\,000$ , nous avons calculé  $\hat{\beta}^{(b)}$  par les moindres carrés généralisés pour la nouvelle méthode GREG ainsi que pour la méthode BLUP. Pour la méthode GREG, nous avons utilisé les moindres carrés ordinaires pour estimer  $\hat{\beta}$  par  $\hat{\beta}_{OLS}^{(b)}$ . En outre, nous avons calculé  $\hat{\Sigma}_v^{(b)}$  et  $\hat{\sigma}_{e(2)}^{(b)}$  par la méthode du maximum de vraisemblance restreint (REML). Pour chaque échantillon produit, nous avons calculé

$$\mu_i^{(b)} = \bar{x}_i'(Z_i'\beta + v_i^{(b)}), \quad i = 1, \dots, m; \quad b = 1, \dots, B.$$

Nous avons calculé le nouvel estimateur GREG de  $\mu_i^{(b)}$  selon  $\hat{\mu}_i^{(b)} = (X_i' - \bar{x}_i')(Z_i'\hat{\beta}^{(b)} + \hat{v}_i^{(b)})$ , l'estimateur GREG de  $\mu_i^{(b)}$  selon  $\hat{\mu}_i^{(b)} = \bar{y}_i' + (X_i' - \bar{x}_i)'\hat{\beta}_{OLS}^{(b)}$  et l'estimateur BLUP empirique (EBLUP) de  $\mu_i^{(b)}$  selon  $\hat{\mu}_i^{(b)} = \hat{\Sigma}_v^{(b)}X_i'\hat{\beta}^{(b)} + \hat{v}_i^{(b)}$ , où  $\hat{v}_i^{(b)} = \hat{\Sigma}_v^{(b)}X_i'\hat{\beta}^{(b)} + \hat{v}_i^{(b)}$ . Ensuite, nous avons calculé la moyenne des erreurs quadratiques moyennes ( $EQM_i$ ) et la moyenne des erreurs relatives absolues ( $ERA_i$ )

$$EQM_i = \frac{1}{m} \sum_{b=1}^m EQM_{i,b} \text{ où } EQM_{i,b} = B^{-1} \sum_{b=1}^B (\hat{\mu}_i^{(b)} - \mu_i^{(b)})^2, \\ ERA_i = \frac{1}{m} \sum_{b=1}^m ERA_{i,b} \text{ où } ERA_{i,b} = B^{-1} \sum_{b=1}^B |\hat{\mu}_i^{(b)} - \mu_i^{(b)}| / \mu_i^{(b)},$$

Les moyennes de domaine de  $x_1$  et  $x_2$  sont calculées d'après l'ensemble complet de données et traitées comme étant les moyennes de population  $\bar{X}_{1i}$  et  $\bar{X}_{2i}$ . Nous avons sélectionné un sous-échantillon aléatoire de 10 % des enregistrements dans chaque petit domaine. La taille globale de l'échantillon est  $n = 3\,876$  et le nombre de petits domaines est  $m = 140$ . En utilisant les valeurs d'échantillon de  $x_1, x_2$  et  $z_i$ , ainsi que les moyennes de population  $\bar{X}_{1i}$  et  $\bar{X}_{2i}$ , nous avons calculé  $EQM(\hat{\mu}_i^G)$ ,  $EQM(\hat{\mu}_i^B)$  et  $EQM(\hat{\mu}_i^{LV})$  en utilisant (8), (9) et (10), respectivement, en traitant les estimations des paramètres de régression,  $\Sigma_v$  et  $\sigma_e^2$  du tableau 1 pour les données complètes comme étant les valeurs réelles. Ensuite, nous avons calculé les valeurs moyennes de l'EQM sur les domaines :

$$EQM^G = \frac{1}{m} \sum_{i=1}^m EQM(\hat{\mu}_i^G), \\ EQM^B = \frac{1}{m} \sum_{i=1}^m EQM(\hat{\mu}_i^B), \\ EQM^{LV} = \frac{1}{m} \sum_{i=1}^m EQM(\hat{\mu}_i^{LV}).$$

Nous définissons l'efficacité relative de  $\hat{\mu}_i^B$  par rapport à  $\hat{\mu}_i^{LV}$  comme étant  $EFF_B$  et l'efficacité relative de  $\hat{\mu}_i^G$  par rapport à  $\hat{\mu}_i^{LV}$  comme étant  $EFF_G$ , où

$$EFF_B = \frac{EQM^B}{EQM^G}, \quad EFF_G = \frac{EQM^G}{EQM^{LV}}.$$

Nous résumons les résultats aux tableaux 2 et 3. Ces derniers révèlent que le nouvel estimateur GREG est un peu plus efficace que l'estimateur GREG habituel en ce qui concerne l'EQM moyenne :  $EFF_{LV} \leq 112\%$ . Toutefois, il est considérablement moins efficace que l'estimateur BLUP sous le modèle à deux niveaux postulé. Par exemple, pour le modèle avec  $z$  et la matrice de covariance diagonale (tableau 2),  $EFF_B = 292\%$  comparativement à  $EFF_{LV} = 106\%$ , et  $EQM_B = 0,62$  comparativement à  $EQM_{LV} = 1,72$ .

**Tableau 2**

Comparaison des estimateurs sur petits domaines : efficacité relative (EFF) et EQM moyenne (EQM) pour le cas de la matrice de covariance diagonale basé sur l'ensemble de données de Moura et Holt (1999)

Mesure de la qualité	Modèle sans z	BLUP	GREG	Nouveau GREG
EFF	100 %	112 %	306 %	292 %
EQM	1,92	1,71	0,62	1,83
			1,72	1,72



et  $x_1$  et  $x_2$  représentent, respectivement, le nombre de pièces et le niveau d'études du chef du ménage (centres chacune autour de leur moyenne de population). Notons que le modèle (12) pour les coefficients  $\beta_j$  aléatoires ne contient pas de covariables  $Z$  au niveau du domaine.

Moura et Holt (1999) ont également étudié un autre modèle avec une covariable  $Z$  au niveau du domaine pour la modélisation des  $\beta_j$  dans (12). Pour ces données, le nombre moyen d'automobiles par ménage dans chaque domaine a été utilisé comme covariable  $z$  pour la modélisation des coefficients aléatoires  $\beta_{j1}$  et  $\beta_{j2}$  correspondant aux variables  $x_1$  et  $x_2$ , mais non pour le terme constant aléatoire,  $\beta_{j0}$ . Le modèle (11) avec la

covariable  $z$  au niveau du domaine est donné par

$$y_{ij} = \beta_{j0} + \beta_{j1}x_{ij1} + \beta_{j2}x_{ij2} + e_{ij}$$

$$j = 1, \dots, N; i = 1, \dots, m \quad (13)$$

avec

$$\beta_{j0} = \beta_0 + v_{j0};$$

$$\beta_{j1} = \beta_1 + \alpha_1 z_i + v_{j1}; \beta_{j2} = \beta_2 + \alpha_2 z_i + v_{j2}. \quad (14)$$

Moura et Holt (1999) ont ajusté les modèles (11)-(12) et (13)-(14) à l'ensemble de données complet susmentionné. Nous résumons leurs résultats au tableau 1.

Tableau 1  
Estimations des paramètres fondées sur l'ensemble de données de Moura et Holt (1999), où  $\sigma_0^2$ ,  $\sigma_1^2$  et  $\sigma_2^2$  sont les éléments diagonaux et  $\sigma_{01}$ ,  $\sigma_{02}$  et  $\sigma_{12}$  sont les éléments non diagonaux de la matrice de covariance  $\Sigma_v$

Paramètre	Covariance	Covariance	Covariance
	diagonale :	diagonale :	générale :
	sans 2	sans 2	avec 2
$\beta_0$	8,442	8,688	8,456
$\beta_1$	0,451	1,321	1,223
$\beta_2$	0,744	2,636	2,596
$\alpha_1$	3,779	-	-
$\alpha_2$	1,659	-	-
$\sigma_0^2$	0,745	0,637	1,385
$\sigma_1^2$	0,237	0,471	0,234
$\sigma_2^2$	0,700	1,472	0,926
$\sigma_{01}$	-	-	0,354
$\sigma_{02}$	-	-	0,492
$\sigma_{12}$	-	-	0,333
$\sigma_e^2$	44,00	44,01	47,74

$E\tilde{Q}M(\mu_B^i) = E(\mu_B^i - \mu_i)^2$   
 $= E[X_i'(\hat{v}_i - v_i)]^2$   
 $= X_i'(\hat{\Sigma}_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v) \bar{X}_i$   
comme il est énoncé dans le théorème 2.

Théorème 2. L'EQM de l'estimateur BLUP (3) est donnée par

$$E\tilde{Q}M(\mu_B^i) = X_i'(\Sigma_v - \Sigma_v X_i' V_i^{-1} X_i \Sigma_v) \bar{X}_i \quad (9)$$

Le théorème 3 donne l'EQM du nouvel estimateur GREG  $\hat{\mu}_{LV}^i$ .

Théorème 3. L'EQM du nouvel estimateur GREG (7) est donnée par

$$E\tilde{Q}M(\mu_{LV}^i) = E\tilde{Q}M(\mu_B^i) + \left\{ \bar{X}_i' \Sigma_v X_i' V_i^{-1} X_i \Sigma_v \bar{X}_i - \bar{X}_i' \Sigma_v \bar{X}_i + \frac{n_i}{\sigma_e^2} \right\}. \quad (10)$$

La preuve du théorème 3 est donnée en annexe.

Par définition, nous avons  $E\tilde{Q}M(\mu_{LV}^i) \leq E\tilde{Q}M(\mu_B^i)$  et (10) donne une expression explicite de l'accroissement de l'EQM de  $\hat{\mu}_{LV}^i$  comparativement à l'EQM de l'estimateur BLUP  $\hat{\mu}_B^i$ .

#### 4. Résultats empiriques

#### 4.1 Comparaison empirique des valeurs de l'EQM

Afin d'étudier l'efficacité du nouvel estimateur GREG, nous avons utilisé des données tirées de Moura et Holt (1999) recueillies auprès de 38 740 ménages dans les districts de recensement (petits domaines) d'un comté du Brésil. Le revenu du chef de ménage a été traité comme la variable de réponse  $y$ . Deux variables indépendantes ont été définies au niveau de l'unité, à savoir le niveau d'études du chef du ménage (échelle ordinaire de 0 à 5) et le nombre de pièces dans le logement du ménage (1 à 11+). Nous avons supposé que le modèle à deux niveaux pour ces données était de la forme :

$$y_{ij} = \beta_{j0} + \beta_{j1}x_{ij1} + \beta_{j2}x_{ij2} + e_{ij} \quad (11)$$
$$j = 1, \dots, N; i = 1, \dots, m \quad (12)$$

$$\beta_{j0} = \beta_0 + v_{j0}; \beta_{j1} = \beta_1 + v_{j1}; \beta_{j2} = \beta_2 + v_{j2}, \quad (12)$$

$$v = (v_{j0}, v_{j1}, v_{j2})' \sim N_3(0, \Sigma_v), e_{ij} \sim N(0, \sigma_e^2) \quad \text{iii}$$

Partant de (5) avec  $\hat{y}_j^* = x_j^{*'} Z_j \beta$  pour prédicteur de  $y_j^*$  sous le modèle (1) avec  $\beta_j^* = Z_j' \beta$  fixes, nous pouvons écrire l'estimateur GREG habituel sous la forme

$$\hat{\mu}_j^* = \bar{y}_j + (\bar{X}_j - \bar{x}_j) Z_j' \beta, \quad (6)$$

où l'indice supérieur  $G$  sur  $\hat{\mu}_j$  signifie GREG (Särndal, Swensson et Wretman 1992, page 225),  $\bar{y}_j$  est la moyenne d'échantillon de  $y_j^*$  dans le  $i^e$  domaine, et  $\bar{x}_j$  est la moyenne d'utilisant dans (5) le prédicteur (4) fondé sur le modèle à deux niveaux (1), nous obtenons un nouvel estimateur GREG de  $\mu_j$  (ou de  $Y_j$ ) de la forme

$$\hat{\mu}_{LV}^* = [\bar{X}_j' (Z_j \beta + \hat{v}_j) + (\bar{y}_j - \bar{x}_j' (Z_j \beta + \hat{v}_j))] \\ = \bar{y}_j + (\bar{X}_j - \bar{x}_j) (Z_j \beta + \hat{v}_j). \quad (7)$$

où l'indice supérieur LV sur  $\hat{\mu}_j$  indique qu'il a été proposé pour la première fois par Lehtonen et Veijanen (1999). Les estimateurs  $\hat{\mu}_{LV}^*$  et  $\hat{\mu}_{LV}^{*c}$  sont linéaires en les  $y_j^*$  et sans biais sous le modèle à deux niveaux (1). En pratique, nous remplaçons les paramètres  $\beta$ ,  $Z_j$  et  $\sigma_e^2$  dans (3), (6) et (7) par des estimateurs appropriés. Les estimateurs résultants sont désignés par  $\hat{\mu}_B^*$ ,  $\hat{\mu}_{LV}^*$  et  $\hat{\mu}_{LV}^{*c}$ , respectivement, où  $\hat{\mu}_B^*$  est l'estimateur BLUP empirique (EBLUP). Sous l'hypothèse de normalité,  $\hat{\mu}_B^*$  est le meilleur estimateur empirique (EB pour *empirical best*). L'estimateur EBLUP de  $Y_j$  est donné à la section 4.2.2. Notons que  $\hat{\mu}_B^*$  et  $\hat{\mu}_{LV}^{*c}$  sont valides en tant qu'estimateurs de  $Y_j$ .

### 3. Erreur quadratique moyenne

L'erreur quadratique moyenne (EQM) de l'estimateur GREG habituel  $\hat{\mu}_j^*$  sous le modèle à deux niveaux peut s'écrire

$$\begin{aligned} EQM(\hat{\mu}_j^*) &= E(\hat{\mu}_j^* - \mu_j)^2 \\ &= E(\hat{\mu}_j^* - \bar{x}_j' Z_j \beta - \bar{X}_j' (Z_j \beta + \hat{v}_j))^2 \\ &= E[(\bar{x}_j - \bar{X}_j)' Z_j \beta + (\bar{X}_j - \bar{x}_j)' \hat{v}_j + e_j]^2 \\ &= (\bar{x}_j - \bar{X}_j)' Z_j \beta (\bar{x}_j - \bar{X}_j) + \frac{\sigma_e^2}{n_j}, \end{aligned}$$

comme il est énoncé dans le théorème 1.

**Théorème 1.** L'EQM de l'estimateur GREG (6) est donnée par

$$EQM(\hat{\mu}_j^*) = (\bar{x}_j - \bar{X}_j)' \Sigma_y (\bar{x}_j - \bar{X}_j) + \frac{\sigma_e^2}{n_j}. \quad (8)$$

En outre, nous pouvons écrire l'EQM de l'estimateur BLUP  $\hat{\mu}_B^*$  comme il suit :

À la section 3, nous commençons par calculer l'erreur

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

Lehtonen, Särndal et Veijanen (2003) ont étudié l'effet du

choix du modèle sur divers types d'estimateurs (synthétique,

GREG et composite) de moyenne de petit domaine.

# Estimation pour petits domaines sous un modèle à deux niveaux

Mahmoud Torabi et J.N.K. Rao<sup>1</sup>

## Résumé

Lehtonen et Veijanen (1999) ont proposé un nouvel estimateur par la régression généralisée (GREG) assisté par modèle d'une moyenne de petit domaine sous un modèle à deux niveaux. Ils ont montré que l'estimateur proposé donne de meilleurs résultats que l'estimateur GREG habituel en ce qui concerne le biais relatif absolu moyen et l'erreur relative absolue médiane moyenne. Nous calculons l'erreur quadratique moyenne (EQM) du nouvel estimateur GREG sous le modèle à deux niveaux et nous la comparons à celle de l'estimateur fondé sur le meilleur prédicteur linéaire sans biais (BLUP). Nous présentons aussi des résultats empiriques concernant l'efficacité relative des estimateurs. Nous montrons que le nouvel estimateur GREG a de meilleures propriétés que l'estimateur GREG habituel en ce qui concerne l'EQM moyenne et l'erreur relative absolue moyenne. Nous montrons aussi que, parce qu'il emprunte de l'information aux petits domaines apparentés, l'estimateur EB-LUP donne des résultats nettement meilleurs que l'estimateur GREG habituel et que le nouvel estimateur GREG. Nous fournissons les résultats de simulation sous un modèle, ainsi qu'en population finie réelle.

Mots clés : Estimateur BLUP, estimateur GREG, erreur quadratique moyenne; effets aléatoires; moyennes de petit domaine.

## 1. Introduction

L'Estimation pour petits domaines a suscité beaucoup d'intérêt ces dernières années à cause de la demande croissante de statistiques fiables pour des petits domaines. Les estimateurs par domaine directs classiques ne sont pas suffisamment précis, parce que, pour les petits domaines, les tailles d'échantillons sont souvent insuffisantes. Il faut donc recourir à des estimateurs indirects renforcés par l'emprunt d'information à des domaines apparentés, en particulier des estimateurs indirects fondés sur un modèle. En estimation sur petits domaines, des modèles à effets aléatoires au niveau de l'unité, dont les modèles de régression linéaire à erreurs emboîtées et les modèles à deux niveaux, sont fréquemment adoptés pour obtenir de bons estimateurs fondés sur un modèle complet de l'Estimation pour petits domaines fondée sur un modèle.

$$y_{ij} = x'_{ij}\beta_j + e_{ij}$$

$$\beta_j = Z_j\beta + v_j, j = 1, \dots, N; i = 1, \dots, m \quad (1)$$

où  $N_j$  est le nombre d'unités dans le  $i^e$  domaine ( $i = 1, \dots, m$ ),  $y_{ij}$  est la réponse et  $x_{ij}$  est un vecteur de dimension  $p \times 1$  de covariables au niveau du domaine,  $\beta$  est un vecteur de dimension  $q \times 1$  de paramètres de régression, les  $v_j$  sont des vecteurs aléatoires indépendants de moyenne nulle et de covariance  $\Sigma_v$ , et les  $e_{ij}$  sont des variables aléatoires indépendantes de moyenne

nulle et de variance  $\sigma_e^2$  qui sont indépendantes des  $v_j$ . Nous pouvons exprimer la moyenne  $\bar{y}_i$  du  $i^e$  domaine sous la forme

$$\bar{y}_i \approx \mu_i = X_i'Z_i'\beta + v_i$$

en supposant que  $N_j$  est grand, où  $\bar{X}_i$  est la moyenne de population connue de  $x_{ij}$  dans le  $i^e$  domaine. Nous supposons que les valeurs d'échantillon  $\{(y_{ij}, x_{ij}), j = 1, \dots, m; i = 1, \dots, m\}$  obéissent au modèle (1), c'est-à-dire qu'il n'y a pas de biais de sélection d'échantillon. Pour l'échantillon, le modèle est donné par

$$y_{ij} = x'_{ij}(Z_j'\beta + v_j) + e_{ij}, j = 1, \dots, m; i = 1, \dots, m. \quad (2)$$

En notation matricielle, (2) peut s'écrire

$$Y_i = X_i(Z_i'\beta + v_i) + e_i, i = 1, \dots, m$$

avec  $\text{Var}(Y_i) = V_i = X_i\Sigma_vX_i' + \sigma_e^2I_m$ , où  $Y_i$  est un vecteur de dimension  $n_i \times 1$  et  $X_i$  est une matrice de dimensions  $n_i \times p$ . Le modèle à deux niveaux (2) a été proposé pour la première fois par Moura et Holt (1999) dans le contexte de l'estimation sur petits domaines. Il intègre efficacement l'utilisation de covariables aux niveaux de l'unité et du domaine dans un seul modèle, par modélisation des pentes aléatoires  $\beta_j$  dans (1) en fonction de covariables au niveau du domaine  $Z_j$ .

Lehtonen et Veijanen (1999) ont proposé un nouvel estimateur par la régression généralisée (GREG) assisté par modèle d'une moyenne de petit domaine sous le modèle à deux niveaux. Ils ont montré que le nouvel estimateur GREG basé sur le modèle (1) donne de meilleurs résultats que l'estimateur GREG habituel fondé sur un modèle avec  $\beta_j = Z_j'\beta$  fixes. Moura et Holt (1999) ont obtenu





C4. Codes R pour la simulation des  $\pi_i$  sous substitution d'unités.

```
p1<-rep(0,N)
for(i in 1:1000000){
  s<-sysppsub(x,n,refus)
  for(j in 1:N){
    if(m1n(abs(s-j))==0) p1[j]<-p1[j]+1
  }
  p1<-p1/1000000
```

## Bibliographie

Asok, C., et Sukharnne, B.V. (1976). On Sampford's procedure of unequal probability sampling without replacement. *Journal of the American Statistical Association*, 71, 912-918.

Chao, M.T. (1982). A general purpose unequal probability sampling plan. *Biometrika*, 69, 653-656.

Chen, X.H., Dempster, A.P. et Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Connor, W.S. (1966). An exact formula for the probability that two specified sampling units occur in a sample drawn with unequal probability and without replacement. *Journal of the American Statistical Association*, 61, 384-390.

Deville, J.C., et Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika*, 85, 89-101.

Fattorini, L. (2006). Applying the Horvitz-Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion probabilities. *Biometrika*, 93, 269-278.

Fellegi, I.P. (1963). Sampling with varying probabilities without replacement: Rotating and non-rotating samples. *Journal of the American Statistical Association*, 58, 183-201.

Goodman, R., et Kish, L. (1950). Controlled selection - A technique in probability sampling. *Journal of the American Statistical Association*, 45, 350-372.

Gray, G.B. (1973). On increasing the sample size (number of psu's). Note technique, Statistique Canada.

Hartley, H.O., et Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33, 350-374.

Hidiroglou, M.A., et Gray, G.B. (1980). Construction of joint probability of selection for systematic P.P.S. sampling. *Applied Statistics*, 29, 107-112.

Kott, P.S. (2005). A note on the Hartley-Rao variance estimator. *Journal of Official Statistics*, 21, 433-439.

Rao, J.N.K. (1965). On two simple schemes of unequal probability sampling without replacement. *Journal Indian Statist. Association*, 3, 173-180.

Sampford, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*, 54, 499-513.

Tillé, Y. (1996). An elimination procedure for unequal probability sampling without replacement. *Biometrika*, 83, 238-241.

ou  $\bar{T} = \sum_{i \in s} y_i^1 / \pi_i^1$ ,  $\bar{T} = \sum_{i \in s} y_i^1 / \pi_i^1$ , et les  $\pi_i^1$  sont les probabilités d'inclusion de premier ordre simulées à partir de  $K$  échantillons indépendants. En notant que  $E(\pi_i^1) = \pi_i$  et  $\text{Var}(\pi_i^1) = \pi_i(1 - \pi_i)/K$ , en vertu de l'ingéatité de Chebyshev, nous avons  $P(|\pi_i^1 - \pi_i| > c) \leq \pi_i(1 - \pi_i)/(Kc^2)$  pour tout  $c > 0$ . Il s'ensuit que :

$$\begin{aligned} P\left(\left|\frac{\pi_i^1}{\pi_i^* - \pi_i}| \leq \varepsilon\right.\right) &= P(\pi_i^1 > \pi_i^* \varepsilon) + P(\pi_i^1 - \pi_i < -\pi_i^* \varepsilon) \\ &= P(\pi_i^1 - \pi_i > \varepsilon \pi_i^* / (1 + \varepsilon)) + P(\pi_i^1 - \pi_i < -\varepsilon \pi_i^* / (1 + \varepsilon)) \\ &\leq P(|\pi_i^1 - \pi_i| > \varepsilon \pi_i^* / (1 + \varepsilon)) + P(|\pi_i^1 - \pi_i| > \varepsilon \pi_i^* / (1 + \varepsilon)) \\ &\leq \frac{K\varepsilon^2\pi_i^2}{(1 + \varepsilon)^2\pi_i(1 - \pi_i)} + \frac{K\varepsilon^2\pi_i^2}{2(1 + \varepsilon^2)\pi_i(1 - \pi_i)} \left(\frac{\pi_i^1}{\pi_i} - 1\right). \end{aligned}$$

Si  $y_i \geq 0$  pour tout  $i$ , alors

$$|\bar{T} - T| \leq \sum_{i \in s} \frac{\pi_i}{y_i} \left| \frac{\pi_i^1}{\pi_i^* - \pi_i} - \frac{\pi_i^1}{\pi_i^*} \right| \leq \max_{i \in s} \left\{ \frac{\pi_i^1}{|\pi_i^* - \pi_i|} \right\} T.$$

Pour tout  $\varepsilon > 0$  et l'échantillon donné  $s$ ,

$$P\left(\left|\frac{\bar{T}}{T} - 1\right| \leq \varepsilon\right) \geq P\left(\max_{i \in s} \left|\frac{\pi_i^1}{\pi_i^* - \pi_i}\right| \leq \frac{\pi_i^1}{\pi_i^*}\right) \geq 1 - \sum_{i \in s} P\left(\left|\frac{\pi_i^1}{\pi_i^*} - \pi_i\right| > \varepsilon\right) \geq 1 - \frac{K\varepsilon^2}{2(1 + \varepsilon^2)} \left(\sum_{i \in s} \frac{\pi_i}{1 - \pi_i}\right).$$

## Annexe C

### Mise en œuvre en R/S-PLUS

C1. Une fonction R pour l'échantillonnage PPT systématique randomisé.  
Les variables d'entrée de la fonction sont  $x$  : le vecteur de population de variables de taille et  $n$  : la taille de l'échantillon. La fonction `sysps` donne l'ensemble de  $n$  unités sélectionnées.

```
sysps<-function(x,n){
  N<-length(x)
  U<-sample(N,n)
  xx<-x[U]
  z<-rep(0,N)
  for(i in 1:N) z[i]<-n*sum(xx[1:i])/sum(x)
  r<-runif(1)
  s<-numeric()
  for(i in 1:N){
    if(z[i]>=x){

```

C2. Une fonction R pour la simulation des probabilités d'inclusion de deuxième ordre.  
Les variables d'entrée de la fonction sont  $x$  : le vecteur de population des variables de taille et  $s$  : l'ensemble d'étiquettes des unités dans l'échantillon. La méthode d'échantillonnage par défaut est l'échantillonnage PPT systématique randomisé et le nombre d'échantillons répétés est  $K = 10^6$ . La fonction `p1j` donne une matrice de dimensions  $n \times n$  où la  $(ij)^{\text{e}}$  entrée est la probabilité simulée  $\pi_{ij}^*$ ,  $i, j \in s$ .

C3. Une fonction R pour l'échantillonnage PPT sous substitution d'unités.  
syspspsub<-function(x,n,refus){  
 s<-sysps(x,n)  
 sub<-numeric()  
 for(i in 1:n){  
 if(min(abs(s[i]-refus))==0) sub<-c(sub,i)  
 }  
 if(min(abs(s[i]-refus))>0) sub<-c(sub,i+1)  
 return(p)  
}  
C3. Une fonction R pour l'échantillonnage PPT sous substitution d'unités.

```
syspspsub<-function(x,n,refus){
  s<-sysps(x,n)
  sub<-numeric()
  for(i in 1:n){
    if(min(abs(s[i]-refus))==0) sub<-c(sub,i)
  }
  if(m>0){
    mc<-length(sub)
    s<-s[sub]
    U1<-1:length(x)[-c(refus,s)]
    x1<-x[-c(refus,s)]
    s1<-sysps(x1,m)
    s<-c(s,U1[s1])
  }
  return(s[order(s)])
}
```

La méthode par défaut pour la sélection de l'échantillon initial et de l'échantillon de substitution est l'échantillonnage PPT systématique randomisé. La fonction `R syspspsub` est utilisée pour simuler les probabilités d'inclusion sous substitution d'unités. Les variables d'entrée sont  $x$  : le vecteur de population des variables de taille,  $n$  : la taille de l'échantillon et `refus` : l'ensemble d'unités de l'échantillon initial qui refusent de répondre. La fonction donne un ensemble d'unités pour l'échantillon final.



Remerciements

Cette étude a été financée en partie par des subventions du Canada. Les auteurs remercient aussi de leur appui les responsables du projet *International Tobacco Control (ITC) Policy Evaluation* et du projet de l'enquête ITC Chine. Le projet ITC est financé en partie par des subventions du Roswell Park Transdisciplinary Tobacco Use Research Center du National Cancer Institute des États-Unis (P50 CA1236) et des Instituts de recherche en santé du Canada (57897). Le financement du projet ITC Chine est assuré par le ministère de la Santé et le ministère des Finances de la Chine.

Annexe A

Un calcul direct sous refus aléatoire

Sous le plan d'échantillonnage PPT systématique randomisé et en supposant que le refus est aléatoire, il est possible, en principe, de calculer directement les probabilités d'inclusion sous une règle de substitution. Le point de départ consiste à énumérer tous les échantillons initiaux possibles et leurs probabilités en se basant sur l'ordre de population particulier utilisé pour sélectionner l'échantillon initial.

Rappelons que  $A_0 = 0$ ,  $A_j = \sum_{i=1}^j (nz_i)$  et  $A_N = n$ . Pour une valeur de départ uniforme choisis  $n \in [0, 1]$ , l'unité  $j$  est sélectionnée si

$$A_{j-1} \leq n + k < A_j \tag{5.1}$$

pour une valeur de  $k = 0, 1, \dots, n-1$ . Soit  $k_j$  le plus grand nombre entier inférieur à  $A_j$ , et soit le reste  $e_j$  donné par  $e_j = A_j - k_j$ . Soit  $0 < e_{(2)} \leq \dots \leq e_{(N)}$  les statistiques d'ordre des restes, et soit  $k_{(1)}, \dots, k_{(N)}$  les valeurs correspondantes de  $k_j$ . Notons que  $e_{(N)} = 1$ . Nous pourrions générer  $N$  échantillons possibles  $s_1^m, \dots, s_N^m$  avec les probabilités respectives

$$e_{(1)}^{(1)}, e_{(2)}^{(2)}, \dots, e_{(N)}^{(N)} - e_{(N-1)}^{(N-1)},$$

dont certaines pourraient être égales à 0. Nous commençons par générer  $s_1^m$ . Partant de chaque unité  $j = 1, \dots, N$ , plaçons  $j$  dans  $s_1^m$  si  $A_{j-1} \leq k < A_j$  pour une valeur  $k = 0, 1, \dots, n-1$ , c'est-à-dire que  $s_1^m$  est sélectionné en utilisant  $n = 0$  dans (5.1). À mesure que nous faisons varier  $n$  de 0 à 1, nous pouvons identifier séquentiellement différents échantillons possibles. Maintenant, sachant que le  $(k_{(m)} + 1)^{\text{e}}$  élément est avancé d'une valeur 1. Par exemple, supposons que  $n = 4$  et  $s^m = \{1, 3, 6, 9\}$ , et supposons que  $k_{(m)} = 0$ , alors  $s^{m+1} = \{2, 3, 6, 9\}$ . Par

ailleurs, si  $k_{(m)} = 2$ , alors  $s^{m+1} = \{1, 3, 7, 9\}$ . L'échantillon  $s^{m+1}$  aura la probabilité  $e_{(m+1)}^{(m+1)} = e_{(m)}^{(m)}$ .

Par construction,  $\pi_i = nz_i$  pour  $i = 1, \dots, N$ . Si l'on souhaite calculer les probabilités d'inclusion de premier et de deuxième ordre seulement, on peut utiliser un algorithme similaire mais simple pour calculer directement les probabilités d'inclusion de deuxième ordre, sachant l'ordre initial. Cependant, dans les applications où les probabilités de tous les échantillons sont nécessaires, l'algorithme de génération d'échantillons peut être exécuté. Ainsi, pour les populations de petite taille, il est alors également possible de calculer les probabilités d'inclusion de premier ordre en cas de refus et de substitution. Supposons que nous sélectionnons pour commencer un échantillon de taille  $n$  par lectionnons pour commencer un échantillon de taille  $n$  par échantillonnage PPT systématique randomisé. Supposons que  $n_1$  unités de cet échantillon consentent à répondre et qu'un nombre  $n_2 = n - n_1$  d'unités supplémentaires soient sélectionnées, de nouveau par échantillonnage PPT systématique randomisé à partir des unités n'ayant pas été échantillonnées la première fois. Supposons pour simplifier que les cas de refus dans le premier échantillon surviennent au hasard et qu'il n'y en a aucun dans l'échantillon de substitution. Notons que cette hypothèse diffère de celle utilisée dans l'exemple 3, où l'ensemble des cas de refus est considéré comme étant non aléatoire. La probabilité d'inclusion de l'unité  $i$ , sachant l'ordre de population initial supposé, est

$$(5.2) \quad nz_i \times \frac{n}{n_1} + \sum_{m: i \in s^m} p_1(s^m) \frac{\sum_{j: j \notin s^m} z_j}{n_2 z_i}.$$

La somme externe est calculée sur l'ensemble des échantillons  $s^m$  de taille  $n$ , générés conformément au processus décrit plus haut, mais sans avoir d'unité  $i$ , avec les probabilités  $p_1(s^m) = e_{(m)}^{(m)} - e_{(m-1)}^{(m-1)}$ . La somme interne qui figure au dénominateur est calculée sur l'ensemble des unités  $j$  non incluses dans les échantillons  $s^m$  de la somme externe. Les probabilités d'inclusion inconditionnelles peuvent être obtenues par calcul de la moyenne appropriée sur l'ensemble des ordres de population qui produisent des valeurs distinctes. Cela n'est manifestement faisable que si la population est petite ou que  $z$  prend un petit nombre de valeurs.

Annexe B

Obtention de l'expression (2.1)

Nous montrons ici que, pour tout  $\varepsilon > 0$  et un échantillon donné  $s$ ,

$$P\left(\left|\frac{\bar{I}}{\bar{I} - \bar{I}_s}\right| \leq \varepsilon\right) \geq 1 - \frac{K\varepsilon^2}{2(1 + \varepsilon^2)} \left(\sum_{i \in s} \frac{\pi_i}{1 - \pi_i}\right),$$

conditionnel et proposé d'évaluer les propriétés de l'estimateur en utilisant les probabilités d'inclusion simulées pour un échantillon donné présentant un intérêt en pratique. La question fondamentale que soulevaient les méthodes basées sur la simulation est celle de la faisabilité des calculs. L'échantillonnage PPT systématique présente un énorme avantage pour ce qui est de l'efficacité des calculs. La méthode d'échantillonnage avec probabilités inégales de Rao-Sampford (Rao 1965; Sampford 1967), par exemple, est une autre méthode d'échantillonnage PPT d'usage répandu. Elle possède plusieurs attributs désirables, comme des expressions explicites pour les probabilités d'inclusion de deuxième ordre, et est plus efficace que l'échantillonnage PPT systématique randomisé (Asok et Sukhtame 1976). Suit une comparaison des temps de processus sous échantillonnage PPT systématique randomisé et sous-échantillonnage PPT de Rao-Sampford pour la simulation des probabilités d'inclusion de premier ordre. La taille d'échantillon est fixée à  $n = 10$  et le nombre d'échantillons simulés est  $K = 10^5$ . Les résultats sont obtenus en utilisant R sur une machine unix à processeur double.

N	PPT systématique	PPT de Rao-Sampford
200	4,7 heures	7,5 heures
100	2,5 heures	5,0 heures
50	1,6 heure	4,4 heures
20	1,2 heure	8,9 heures

Il convient de souligner que, si la méthode de Rao-Sampford s'appuie sur une procédure réjective et qu'il faut où  $N = 20$ . Cela tient au fait que la méthode de Rao-Sampford s'appuie sur un grand nombre de rejets pour arriver à un échantillon final lorsque la fraction d'échantillonnage  $n/N$  est grande. L'échantillonnage PPT systématique randomisé, quant à lui, n'est pas affecté par cette situation, et la méthode basée sur la simulation peut fournir rapidement des résultats ayant l'exactitude souhaitée pour  $N = 400$ , voire une valeur plus élevée. Plusieurs fonctions R/S-PLUS et les principaux codes pour l'approche proposée sont présentés à l'annexe C et sont applicables à d'autres scénarios de substitution moyennant de légères modifications.

L'une des raisons de l'utilisation de l'échantillonnage PPT systématique randomisé pour sélectionner les grappes de niveau supérieur dans l'enquête ITC Chine est que le plan final est autopondéré. Un problème se pose en cas de refus de participer. Comment faut-il sélectionner les unités de substitution de façon que le plan d'échantillonnage modifié final soit encore (approximativement) autopondéré? Dans d'autres circonstances, comme l'échantillonnage avec renouvellement, cela est réalisable; consulter, par exemple, Fellegi (1963). La façon d'atteindre cet objectif dans le cas du plan de l'enquête ITC Chine est à l'étude.

#### 4. Quelques remarques supplémentaires

L'approximation de Hartley-Rao  $\pi_{ij}$  ayant le biais le plus important. Dans le cas des estimateurs de variance utilisant les  $\pi_{ij}^*$  simulées, l'accroissement de la valeur de  $K$  de  $10^5$  (c'est-à-dire  $v_2$ ) à  $10^6$  (c'est-à-dire  $v_3$ ) rend le biais négligeable, quoique celui obtenu pour  $K = 10^5$  soit clairement acceptable en pratique. Pour ce qui est de l'instabilité, les mesures sont comparables pour les trois versions de l'estimateur de la variance.

En théorie, la méthode de calcul des probabilités d'inclusion basée sur la simulation s'applique à n'importe quel plan d'échantillonnage, à condition que l'on dispose de renseignements complets sur ce dernier. Elle convient bien au traitement de scénarios de substitution plus complexes ou d'autres types de modification du plan original. Dans le cas de l'enquête ITC Chine, l'une des unités ayant refusé de participer a dû être remplacée par une unité provenant d'une région particulière de la ville, à cause de contraintes de charge de travail et de restrictions relatives au travail sur le terrain. Dans celui d'une enquête nationale auprès des jeunes réalisée au Canada, certaines unités ont refusé de participer au deuxième, ainsi qu'au troisième degré (écoles) et des unités de substitution ont donc été sélectionnées avant d'atteindre la taille d'échantillon cible. Comme l'a fait remarquer un rédacteur associé, une situation comparable a également été mentionnée lors du 57<sup>e</sup> cycle de l'enquête de la National Sample Survey Organization du gouvernement de l'Inde ([www.mosp.gov.in](http://www.mosp.gov.in)), où une modification avait été apportée à l'échantillonnage systématique circulaire avec probabilité proportionnelle à la taille afin de sélectionner deux sous-échantillons distincts. Gray (1973) a décrit une méthode d'accroissement de la taille de l'échantillon (nombre d'UPB) quand l'échantillon initial a été sélectionné par la méthode d'échantillonnage PPT systématique randomisée. Le calcul des probabilités d'inclusion de deuxième ordre sous les méthodes proposées est difficile, même si l'échantillon est de très petite taille. Dans tous ces cas, les solutions analytiques pour les probabilités d'inclusion sont difficiles à utiliser ou n'existent pas, mais l'approche fondée sur la simulation peut être appliquée sans trop de difficulté.

Dans un article récent, Fattorini (2006) discute de l'utilisation de la méthode basée sur la simulation dans le cas de l'échantillonnage spatial où les unités sont sélectionnées séquentiellement. Quand un plan d'échantillonnage PPT est atteint par une ou plusieurs séries de substitutions d'unités, le plan modifié peut également être considéré comme séquentiel. Toutefois, nos résultats théoriques sur l'exactitude des méthodes basées sur la simulation diffèrent de ceux de Fattorini. Nous avons utilisé un argument

3.2 Probabilités d'inclusion de deuxième ordre

L'échantillonnage PPT systématique randomisé a suscité de nombreux travaux de recherche, principalement en vue d'obtenir les probabilités d'inclusion de deuxième ordre  $\pi_{ij}^*$  et les estimateurs de variance. Hartley et Rao (1962) ont établi des formules exactes pour les  $\pi_{ij}^*$  quand  $n = 2$  et  $N = 3$  ou  $N = 4$ ; Connor (1966) a étendu les résultats et dérivé la formule exacte pour les cas généraux  $n$  et  $N$ , et la méthode de calcul commence à être implémentée plus tard dans le langage Fortran par Hidiroglou et Gray (1980). La procédure est assez lourde, comme en témoignent les 165 lignes de code Fortran.

Le résultat le plus curieux est probablement l'approximation asymptotique des  $\pi_{ij}^*$  calculées par Hartley et Rao (1962). Dans un papier récent, Kott (2005) a montré que l'estimateur de variance d'un estimateur d'Horvitz-Thompson fondé sur l'approximation de Hartley-Rao non seulement donne de bons résultats sous le cadre fondé sur le plan de sondage, mais a aussi de bonnes propriétés sous le modèle. L'approximation de Hartley-Rao a été dérivée au départ en partant de l'hypothèse que  $n$  est fixe et que  $N$  est grand et correct jusqu'à l'ordre  $O(N^{-4})$  (Hartley et Rao 1962 : Équation (5.15) à la page 369). Lors d'une conversation privée avec J.N.K. Rao durant le 23<sup>e</sup> Symposium international sur les questions de méthodologie de Statistique Canada, celui-ci a fait remarquer que l'approximation est encore valide, même si  $n$  est grand, à condition que  $n/N$  soit faible. Dans les cas où  $N$  n'est pas grand (et/ou) que  $n/N$  n'est pas faible, comme dans l'exemple de l'enquête ITC Chine considéré ici, aucune donnée n'a été publiée au sujet de l'approximation de Hartley-Rao.

Quand la méthode d'échantillonnage PPT systématique randomisé est altérée par la substitution d'unités, il est virtuellement impossible de calculer les probabilités d'inclusion de deuxième ordre ou toute approximation. En revanche, selon l'approche basée sur la simulation, l'obtention d'estimations très fiables des  $\pi_{ij}^*$  à l'aide d'un grand nombre d'échantillons simulés reste simple, à condition que la procédure d'échantillonnage altérée soit entièrement spécifiée. Nous examinons dans la suite les propriétés des estimateurs de variance en utilisant les  $\pi_{ij}^*$  simulées lorsque la méthode d'échantillonnage PPT systématique randomisé n'est pas altérée. Dans ces conditions,  $\pi_i = n_i$  et l'approximation  $\pi_{ij}^*$  de Hartley-Rao pour  $\pi_{ij}^*$  peut aussi être incluse dans la comparaison.

*Exemple 4.* Nous commençons par comparer  $\pi_{ij}^*$  à  $\pi_{ij}$  pour chacune des entrées individuelles. Faute d'espace, nous présentons uniquement les résultats pour  $i = 1, \dots, 5$  et  $j = 1, \dots, 10$ , qui suffisent à dépendre le tableau général. L'approximation de Hartley-Rao  $\pi_{ij}^*$  est très proche des  $\pi_{ij}$  simulées, la concordance allant jusqu'à la deuxième

décimale pour la majorité des entrées. Il s'agit manifestement d'une observation intéressante, sachant que  $N = 20$  et  $n = 10$ .

i)  $\pi_{ij}^*$  simulées,  $K = 10^5$  :

0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873
0,0000	0,3122	0,3821	0,2975	0,1669	0,1442	0,2116	0,2249	0,3975	0,1873

ii) Approximation de Hartley-Rao  $\pi_{ij}^*$  :

0,0000	0,3079	0,3769	0,2932	0,1669	0,1427	0,2143	0,2286	0,3921	0,1884
0,3079	0,0000	0,3569	0,0000	0,1351	0,2029	0,2164	0,3712	0,1784	0,2193
0,3769	0,3569	0,0000	0,3421	0,1296	0,1946	0,2075	0,3559	0,1710	0,2292
0,2932	0,0000	0,3421	0,0000	0,1514	0,2132	0,1946	0,2075	0,3559	0,1710
0,1669	0,1351	0,1296	0,1514	0,0000	0,0732	0,1109	0,2010	0,0966	0,1168

*Exemple 5.* Dans le cas des probabilités d'inclusion de deuxième ordre, l'aspect principal est l'estimation de la variance. Pour une taille d'échantillon fixe, un estimateur sans biais de la variance de l'estimateur d'Horvitz-Thompson  $\hat{Y}_{HT} = \sum_{i \in s} y_i / \pi_i$  est donné par la forme bien connue de Yates-Grundy,

$$V(\hat{Y}_{HT}) = \sum_{i=1}^n \sum_{j=1}^{f-i+1} \frac{\pi_{ij}}{\pi_i \pi_j - \pi_{ij}} \left( \frac{\pi_i}{y_i} - \frac{\pi_j}{y_j} \right)^2. \quad (3.1)$$

Considérons les trois populations synthétiques décrites dans l'exemple 2. Nous obtenons la variance réelle  $V = \text{Var}(\hat{Y}_{HT})$  par simulation en utilisant  $B = 10^5$  échantillons simulés et en la calculant sous la forme  $B^{-1} \sum_{b=1}^B (\hat{Y}_b - \bar{Y})^2$ , où  $\bar{Y}$  est le total de population réel et  $\hat{Y}_b$  est l'estimateur d'Horvitz-Thompson de  $Y$  calculé d'après le  $b^{\text{e}}$  échantillon simulé. Nous examinons trois estimateurs de la variance de la forme (3.1), désignés respectivement par  $v_1, v_2$  et  $v_3$ , en remplaçant les  $\pi_{ij}$  dans (3.1) respectivement par les  $\pi_{ij}^*$  simulées pour  $K = 10^5$  et les  $\pi_{ij}^*$  pour  $K = 10^6$ . Nous évaluons la performance de ces estimateurs en nous appuyant sur le biais relatif simulé BR =  $(B^{-1} \sum_{b=1}^B v^{(b)} - V) / V$  et l'instabilité simulée INST =  $(B^{-1} \sum_{b=1}^B v^{(b)})^2 / V$ , où  $v^{(b)}$  est l'estimation de la variance calculée d'après le  $b^{\text{e}}$  échantillon, en utilisant un autre ensemble de  $B = 10^5$  échantillons indépendants. Les résultats sont résumés au tableau 1 ci-après. Les trois populations sont indiquées par le coefficient de corrélation  $\rho$  entre  $y$  et  $x$ .

Tableau 1 Biais relatif et instabilité des estimateurs de variance

Population	BR (%)			INST		
	$v_1$	$v_2$	$v_3$	$v_1$	$v_2$	$v_3$
$\rho = 0,30$	6,1 %	1,4 %	-0,3 %	0,66	0,65	0,65
$\rho = 0,50$	4,3 %	2,5 %	-1,1 %	0,42	0,44	0,42
$\rho = 0,80$	2,6 %	1,2 %	-0,2 %	0,61	0,60	0,60

En ce qui concerne le biais relatif, les trois estimateurs de variance sont acceptables, celui ( $v_1$ ) basé sur



d'observation de l'enquête varie de  $N = 20$  à  $N = 120$ . Dans chaque ville,  $n = 10$  districts sont sélectionnés par la méthode d'échantillonnage PPT systématique randomisé. En cas de refus, des districts de substitution sont sélectionnés parmi ceux non inclus dans l'échantillon initial, de nouveau par échantillonnage PPT systématique randomisé. Pour les besoins de l'illustration, nous utilisons l'information sur le plan de sondage provenant de la plus petite ville (c'est-à-dire  $N = 20$ ). Des commentaires supplémentaires sur les cas pour lesquels  $N$  est grand sont formulés à la section 4.

3.1 Probabilités d'inclusion de premier ordre

Nous démontrons d'abord l'exactitude des valeurs simulées de  $\pi_i$  lorsque les valeurs exactes sont connues. Puis, nous examinons l'effet des substitutions d'unités sur la valeur finale de  $\pi_i$  et les propriétés de l'estimateur d'Horvitz-Thompson pour un total de population en utilisant les valeurs simulées de  $\pi_i$ . Les probabilités d'inclusion simulées en présence de substitution d'unités sont comparées à celles obtenues en supposant que le plan modifié est encore un échantillonnage PPT.

*Exemple 1.*  $\pi_i^*$  basées sur la simulation en l'absence de refus. Dans ce cas, les valeurs exactes de  $\pi_i$  sont données par  $\pi_i = n z_i$ .

i) Valeurs exactes de  $\pi_i$  :

0,5840 0,5947 0,6702 0,5331 0,3085 0,2652 0,3930 0,4180 0,6952 0,3471  
0,5993 0,5953 0,8240 0,6868 0,4469 0,2191 0,4237 0,4180 0,7567 0,3163

ii)  $\pi_i^*$  simulées,  $K = 10^5$  :

0,5828 0,5945 0,6656 0,5339 0,3071 0,2656 0,3929 0,4205 0,6969 0,3474  
0,5999 0,5929 0,8227 0,6865 0,4446 0,2186 0,4215 0,4179 0,7569 0,3194

iii)  $\pi_i^*$  simulées,  $K = 10^6$  :

0,5836 0,5955 0,6701 0,5336 0,3081 0,2654 0,3931 0,4180 0,6950 0,3466  
0,5994 0,5994 0,8242 0,6864 0,4469 0,2186 0,4237 0,4172 0,7569 0,3166

Les  $\pi_i^*$  simulées concordent avec les  $\pi_i$  jusqu'à la deuxième décimale pour  $K = 10^5$  et jusqu'à la troisième pour  $K = 10^6$  dans la plupart des cas.

*Exemple 2.* Afin d'évaluer les propriétés de l'estimateur d'Horvitz-Thompson (HT) pour un total de population en utilisant les valeurs réelles de  $\pi_i$  et les  $\pi_i^*$  simulées de l'exemple 1, nous avons généré la variable de réponse  $y_i' = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, \dots, N$ , où  $x_i$  est la variable de taille et les  $\varepsilon_i$  sont indépendants et suivent la même loi normale de moyenne 0 et de variance  $\sigma^2$ . Nous avons considéré trois populations (trois valeurs de  $\sigma^2$ ) pour lesquelles les coefficients de corrélation de population entre  $x$  et  $y$  sont, respectivement, 0,3, 0,5 et 0,8. Pour chacune des trois populations, nous avons sélectionné  $B = 2000$  échantillons répétés de taille  $n = 10$  par échantillonnage PPT systématique randomisé, et pour chaque échantillon, nous avons calculé trois estimateurs HT en utilisant les

valeurs réelles de  $\pi_i$ , les  $\pi_i^*$  simulées avec  $K = 10^5$  et les  $\pi_i^*$  simulées avec  $K = 10^6$ , respectivement. Les résultats, qui ne sont pas présentés ici faute d'espace, montrent que le biais relatif de chacun des trois estimateurs HT est inférieur à 0,04 % et que leurs erreurs quadratiques moyennes sont presque identiques.

*Exemple 3.* Lorsque des refus de participation se produisent dans l'échantillon PPT initial et que des unités de substitution sont sélectionnées parmi les unités non incluses dans l'échantillon initial en utilisant la même méthode d'échantillonnage PPT, deux questions se posent : (1) comment calculer les probabilités d'inclusion  $\pi_i$  pour l'échantillon final et (2) dans quelle mesure la méthode de substitution a-t-elle altéré le plan d'échantillonnage PPT original? Nous pouvons calculer les  $\pi_i^*$  simulées et les comparer aux  $\pi_i^*$  obtenues en supposant que l'échantillonnage demeure PPT après que l'on ait retiré de la base de sondage les unités ayant exprimé un refus. En simulant les  $\pi_i^*$ , nous supposons pour simplifier qu'aucun refus n'est possible parmi les unités non comprises dans l'échantillon initial et, par conséquent, qu'il n'existe aucun refus parmi les unités de substitution. Le nombre de répétitions  $K$  est fixé à  $10^6$  pour la simulation. Nous envisageons deux scénarios dans les conditions où la population contient trois unités qui refusent de participer et que toutes ces unités sont sélectionnées dans l'échantillon initial de taille  $n = 10$ .

- i) Refus de trois grandes unités :  $\pi_i^*$  simulées (deux premières lignes) contre  $\pi_i$  (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT.
- ii) Refus de trois petites unités :  $\pi_i^*$  simulées (deux premières lignes) contre  $\pi_i$  (deux dernières lignes) sous l'hypothèse d'un échantillonnage PPT.

0,6326 0,6049 0,7167 0,5829 0,0000 0,0000 0,4415 0,4666 0,7406 0,3937  
0,6482 0,5901 0,8558 0,7330 0,4965 0,0000 0,4728 0,4664 0,7976 0,3930  
0,6343 0,6025 0,7280 0,5790 0,0000 0,0000 0,4269 0,4540 0,7550 0,3770  
0,6510 0,5858 0,7459 0,7459 0,4854 0,0000 0,4402 0,4540 0,8218 0,3436

Il est évident que la taille des unités qui refusent a un effet très important sur la distribution des probabilités d'inclusion finales. Si l'on omet de tenir compte de l'altération du plan d'échantillonnage résultant de la substitution des unités et que l'on traite ce plan comme s'il s'agissait encore d'un échantillonnage PPT, les probabilités d'inclusion pour les grandes unités sont exagérées et le rôle des petites unités est atténué. Cette tendance est plus prononcée lorsque de grandes unités figurent parmi celles qui refusent de participer, c'est-à-dire le cas (i) où  $\pi_i^* = 0,8070$  comparativement à  $\pi_i = 0,8897$  et  $\pi_i^* = 0,3210$  comparativement à  $\pi_i = 0,2837$ .

d'échantillonnage  $p_i$ , soit  $M_i$  le nombre d'échantillons qui contiennent l'unité  $i$ . Alors, la probabilité d'inclusion de premier ordre  $\pi_i = P(i \in s)$  peut être estimée par  $\pi_i^* = M_i/K$ . Pour un  $i$  particulier, les  $M_i$  suivent une loi binomiale et les  $\pi_i^*$  satisfont  $E(\pi_i^*) = \pi_i$  et  $\text{Var}(\pi_i^*) \leq (4K)^{-1}$ . Si, par exemple, nous pouvons nous permettre de prendre  $K$  aussi grand que  $25 \times 10^6$ , alors,  $P(|\pi_i^* - \pi_i| < 0,001) \geq 0,99$  pour toute probabilité donnée  $\pi_i$ .

Une mesure plus pertinente de l'exactitude des méthodes basées sur la simulation est la performance de l'estimateur d'Horvitz-Thompson lorsqu'on utilise les probabilités d'inclusion simulées. Soit  $\bar{T} = \sum_{i \in s} y_i / \pi_i$  et  $\bar{T}^* = \sum_{i \in s} y_i^* / \pi_i^*$ . Pour un échantillon donné, le biais relatif dû à l'utilisation de  $\bar{T}^*$  au lieu de  $\bar{T}$  est défini par  $(\bar{T}^* - \bar{T}) / \bar{T}$ . Sans perte de généralité, nous supposons que  $y_i \geq 0$  pour tout  $i$ . Nous montrons à l'annexe B que, pour tout  $\varepsilon > 0$  et l'échantillon donné  $s$ ,

$$P\left(\frac{\bar{T}^*}{\bar{T}} - 1 \leq \varepsilon\right) \geq 1 - \frac{K\varepsilon^2}{2(1+\varepsilon^2)} \left(\sum_{i \in s} \frac{\pi_i}{1 - \pi_i} - n\right). \quad (2.1)$$

Comme  $\sum_{i \in s} (1/\pi_i)$  est l'estimateur d'Horvitz-Thompson de la population de taille  $N$ , une borne inférieure pratique pour  $P(|\bar{T}^* - \bar{T}|/\bar{T} \leq \varepsilon)$  avec un petit  $\varepsilon$  est donnée par

$$\Delta = 1 - \frac{K\varepsilon^2}{2(N-n)}. \quad (2.2)$$

Si l'on veut que  $\varepsilon = 0,01$  et  $\Delta = 0,98$ , alors pour  $N - n = 100$  le nombre (théorique) d'échantillons indépendants requis pour la simulation est  $K = 10^8$ . Puisque la borne inférieure donnée par (2.1) est prudente, et valide pour toute variable de réponse, on attendrait à ce qu'un  $K$  plus petit ayant une valeur autour de  $10^7$ , voire même  $10^6$ , donne de bons résultats pour la plupart des scénarios pratiques où  $N - n \leq 100$ , ce que corroborent les exemples numériques présentés à la section 3.

L'estimation des probabilités d'inclusion de deuxième ordre  $\pi_{ij} = P(i, j \in s)$  n'impose aucune difficulté supplémentaire, excepté que le nombre total d'échantillons simulés,  $K$ , requis pour obtenir le même niveau d'exactitude relative que pour les probabilités de premier ordre est plus élevé. Soit  $M_{ij}$  le nombre d'échantillons simulés parmi les  $K$  échantillons indépendants qui contiennent à la fois  $i$  et  $j$ . Soit  $\pi_{ij}^* = M_{ij}/K$  l'estimation pour  $\pi_{ij}$ . Supposons que le but soit d'estimer une grandeur de population quadratique

$$\bar{Q} = \sum_{i=1}^I \sum_{j=1}^I q(V_i, V_j),$$

Les estimateurs de type Horvitz-Thompson de  $\bar{Q}$  en utilisant  $\pi_{ij}^*$  ou  $\pi_{ij}^*$  sont donnés, respectivement, par

$$\bar{\hat{Q}} = \sum_{i \in s} \sum_{j \in s} q(V_i, V_j) \frac{\pi_{ij}}{\pi_i^* \pi_j^*} \quad \text{et} \quad \bar{\tilde{Q}} = \sum_{i \in s} \sum_{j \in s} q(V_i, V_j) \frac{\pi_{ij}}{\pi_{ij}^*}.$$

En suivant le même argument que celui menant à (2.1), nous pouvons montrer que

$$P\left(\left|\frac{\bar{\tilde{Q}}}{\bar{Q}} - 1\right| \leq \varepsilon\right) \geq 1 - \frac{K\varepsilon^2}{2(1+\varepsilon^2)} \left(\sum_{i \in s} \sum_{j \in s} \frac{\pi_{ij}}{1 - \pi_{ij}^*} - n^2\right). \quad (2.3)$$

Comme  $\sum_{i \in s} \sum_{j \in s} (1/\pi_{ij})$  est un estimateur de  $N^2$  sans biais sous le plan, une borne inférieure pratique pour  $P(|\bar{\tilde{Q}} - \bar{Q}|/\bar{Q} \leq \varepsilon)$  est donnée par  $1 - 2(N + n)(N - n)/K^2$ . Si nous la comparons à  $\Delta$  donnée par (2.2), il est évident que nous avons besoin d'un  $K$  beaucoup plus grand pour obtenir la même borne inférieure, quoique, dans les deux cas, la borne inférieure est prudente et la valeur de  $K$  effectivement requise peut être plus faible. Par ailleurs, les probabilités d'inclusion de deuxième ordre sont utilisées pour estimer les paramètres de deuxième ordre, tels que la variance de la population ou la variance d'un estimateur linéaire. L'exactitude souhaitée de l'estimation est moins cruciale que dans le cas des paramètres de premier ordre, comme le total ou la moyenne de population et, par conséquent, un nombre compris entre  $10^6$  et  $10^7$  devrait être acceptable pour  $K$  dans de nombreuses situations pratiques.

L'aspect le plus critique pour les méthodes fondées sur la simulation est manifestement la faisabilité des calculs. Entre

autres, elle dépend en grande partie de la valeur de  $K$  choisie, de la complexité du plan d'échantillonnage et de la puissance de calcul disponible. Si  $K = 10^6$  et que l'on souhaite obtenir les résultats basés sur la simulation en dix heures, il faut tirer 28 échantillons simulés par seconde. L'échantillonnage PPT systématique randomisé est la méthode d'échantillonnage avec probabilités inégales sans remise la plus efficace en ce qui concerne l'exécution des calculs. Elle ne comporte qu'un classement aléatoire simple et la sélection d'un point de départ aléatoire. La plupart des méthodes concurrentes sont des méthodes d'échantillonnage réjectif ou des sélections séquentielles compliquées pour lesquelles le tirage d'échantillons simulés est beaucoup plus long. Nous comparons à la section 4 les temps de processus requis pour calculer les  $\pi_i$  simulés pour l'échantillonnage PPT systématique randomisé et pour l'échantillonnage avec probabilités inégales de Rao-Sampford.

### 3. Exemples numériques

L'information sur le plan de sondage utilisée à la présente section est adaptée d'enquête ITC Chine. Le nombre de districts de voirie (grappes de niveau supérieur) dans chacune des sept villes faisant partie du champ



Bien que le choix d'une méthode d'échantillonnage PPT puisse être souhaitable d'un point de vue théorique, l'exécution est souvent compliquée et parfois même impossible, en raison de contraintes et de limites pratiques. Certains compromis et modifications sont nécessaires. Toutefois, le plan modifié ne satisfait plus la condition (1.1). Le calcul direct des probabilités finales d'inclusion devient souvent difficile, voire impossible. Parmi les problèmes courants qui, dans la pratique des sondages, nécessitent l'altération du plan d'échantillonnage original, les plus fréquents sont les unités qui refusent de participer et la substitution d'unités. L'exemple qui suit illustre cette situation.

L'enquête réalisée par la Chine dans le cadre du projet *International Tobacco Control (ITC) Policy Evaluation Survey* (Enquête ITC Chine) repose sur un plan d'échantillonnage à plusieurs degrés avec probabilités inégales pour la sélection de fumeurs et de non-fumeurs adultes dans sept villes. Chaque ville possède une structure administrative hiérarchique naturelle

Ville → District de voisin → Ilot résidentiel → Ménage → Individu

L'enquête réalisée par la Chine dans le cadre du projet *International Tobacco Control (ITC) Policy Evaluation Survey* (Enquête ITC Chine) repose sur un plan d'échantillonnage à plusieurs degrés avec probabilités inégales pour la sélection de fumeurs et de non-fumeurs adultes dans sept villes. Chaque ville possède une structure administrative hiérarchique naturelle.

Individu → Ménage → Ilot résidentiel → District de voirie → Ville

Aux niveaux supérieurs, la méthode d'échantillonnage PPT systématique randomisée est appliquée pour sélectionner dix districts de votre ville, avec probabilité proportionnelle à la taille de la population du district, puis deux lots résidentiels sont tirés dans chaque district sélectionné, de nouveau par échantillonnage PPT systématique randomisée avec probabilité proportionnelle à la taille des ménages et les individus dans les ménages sont ensuite sélectionnés en utilisant une méthode d'échantillonnage aléatoire simple modifiée. Le plan original consistait à sélectionner 40 fumeurs adultes et 10 non-fumeurs adultes dans chacun des 20 lots résidentiels, afin d'obtenir un échantillon final de 800 fumeurs et de 200 non-fumeurs dans chaque ville.

et de 200 non-fumeurs dans chaque ville. Cependant, un problème s'est posé durant l'exécution de l'enquête : plusieurs grappes de niveau supérieur

Cependant, un problème s'est posé durant l'exécution de l'enquête : plusieurs grappes de niveau supérieur sélectionnées (d'abord les districts de voirie, puis les îlots résidentiels) ont refusé de participer à l'enquête, à cause de conflits avec d'autres activités ou de la non-disponibilité des ressources humaines. Ces grappes qui refusent doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial, une possibilité qui consiste à recourir de nouveau à l'échantillonnage PPT systématique randomisé pour obtenir la taille d'échantillon global cible.

Sous les plans d'échantillonnage à plusieurs degrés tels que celui utilisé pour l'enquête ITC Chine, les probabilités d'inclusion de premier ordre des individus sélectionnés dans l'échantillon final peuvent être calculées en multipliant les

## 2. Propriétés des méthodes basées sur la simulation

Dans le présent article, nous démontrons, en nous appuyant sur des arguments théoriques et des exemples numériques, que les probabilités d'inclusion de premier et de deuxième ordre peuvent être estimées exactement par des simulations de Monte Carlo lorsque l'on dispose de renseignements complets sur le plan de sondage. Nos exemples numériques sont motivés par l'enquête ITC Chine pour laquelle l'échantillonnage PPT systématique randomisé sert de méthode de référence, mais nos résultats à d'autres méthodes d'échantillonnage avec probabilités inégales sans remise. À la section 2, nous présentons les simulations relatives à l'exactitude des méthodes fondées sur la simulation. À la section 3, nous donnons des exemples numériques et des comparaisons. À l'annexe C, nous présentons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée, qui ont été élaborés au départ pour l'enquête ITC Chine. À la section 4, nous formulons certaines remarques supplémentaires.

probabilités d'inclusion des unités aux diverses étapes. Lorsque la méthode d'échantillonnage PPT systématique randomisée est modifiée à cause de la substitution d'unités à une certaine étape, la condition (1.1) n'est plus vérifiée pour l'échantillon final à cette étape. Le calcul des probabilités d'inclusion de premier ordre devient alors très difficile et celui des probabilités d'inclusion de deuxième ordre, virtuellement impossible. À l'annexe A, nous donnons une méthode de calcul direct (5.2) des  $\pi_j$  quand l'échantillon initial et l'échantillon de substitution sont sélectionnés tous deux par échantillonnage PPT systématique randomisé, en supposant que le refus de participer est aléatoire dans l'échantillon initial et qu'aucun refus n'a lieu dans l'échantillon de substitution. L'expression est valide de conditionnellement au nombre de refus et à l'ordre de population utilisé (après randomisation) pour la sélection de l'échantillon initial. Il est évident que, même sous des conditions et hypothèses aussi contraignantes, le traitement de l'expression proprement dite devient difficile pour une taille d'échantillon qui n'est pas tellement grande.

Si le calcul des probabilités d'inclusion exactes est impossible ou prohibitif, mais que l'on dispose de renseignements complets sur le plan de sondage, il est facile d'utiliser des méthodes de simulation de Monte Carlo pour estimer les probabilités d'inclusion. Désignons le plan d'échantillonnage probabiliste entièrement spécifié par  $p$ . La méthode basée sur la simulation est simple, à savoir tirer  $K$  échantillons indépendants, tous selon le même plan



# Echantillonnage PPT systématique basé sur la simulation en cas de substitution d'unités

Mary E. Thompson et Changbao Wu<sup>1</sup>

## Résumé

L'enquête réalisée par la Chine dans le cadre du projet *International Tobacco Control (ITC) Policy Evaluation Survey* comprend un plan d'échantillonnage à plusieurs degrés avec probabilités inégales où les grappes du niveau supérieur sont sélectionnées par la méthode d'échantillonnage PPT systématique randomisée. Durant l'exécution de l'enquête, il faut résoudre le problème suivant : plusieurs grappes de niveau supérieur échantillonnées refusent de participer à l'enquête et doivent être remplacées par des unités de substitution sélectionnées parmi les unités non incluses dans l'échantillon initial, de nouveau par la méthode d'échantillonnage PPT systématique randomisée. Dans de telles conditions, les probabilités d'inclusion de premier ordre des unités finales sélectionnées sont très difficiles à calculer et la détermination des probabilités d'inclusion de deuxième ordre devient pratiquement impossible. Dans le présent article, nous élaborons une méthode fondée sur la simulation pour calculer les probabilités d'inclusion de premier et de deuxième ordre lorsque le calcul direct est prohibitif ou impossible. Nous démontrons l'efficacité et la faisabilité de la méthode que nous proposons en nous appuyant sur des considérations théoriques et des exemples numériques. Nous incluons plusieurs fonctions et codes R/S-PLUS pour la procédure proposée. La méthode peut être étendue à des situations de refus/substitution plus complexes susceptibles de survenir en pratique.

Mots clés : Probabilité d'inclusion; estimateur d'Horvitz-Thompson; méthode de Rao-Sampford; biais relatif; échantillonnage avec probabilités inégales sans remise.

## 1. Introduction

Dans l'analyse de données d'enquête complexe, la construction des poids de sondage est la première étape critique. Elle débute par le calcul des probabilités d'inclusion de premier ordre, qui est souvent simple si le plan d'échantillonnage original est bien exécuté, sans aucune altération ni modification. Par exemple, si les unités d'échantillonnage sont sélectionnées avec probabilité d'inclusion ( $\pi$ ) proportionnelle à la taille (PPT ou  $\pi p_i$ ), les probabilités d'inclusion s'obtiennent facilement par un simple rééchantillonnage de la variable de taille. Parmi les méthodes d'échantillonnage PPT avec probabilités inégales sans remise utilisables lorsque la taille d'échantillon est fixée arbitrairement, la méthode d'échantillonnage PPT systématique randomisée est la plus simple à appliquer. Elle a été décrite pour la première fois dans Goodman et Kish (1950) à titre de méthode de sélection contrôlée, et a été perfectionnée par Hartley et Rao (1962) qui ont étudié le problème important, mais difficile, du calcul des probabilités d'inclusion de deuxième ordre. Soit  $x_i, i = 1, 2, \dots, N$  les valeurs connues de la variable de taille, où  $N$  est le nombre total d'unités dans la population. Soit  $z_i = x_i / X$ , où  $X = \sum_{i=1}^N x_i$ , et supposons que  $n z_i < 1$  pour tout  $i$ . La méthode d'échantillonnage PPT systématique randomisée procède comme suit. Arranger les  $N$  unités de la population dans un ordre aléatoire et poser que  $A_0 = 0$  et  $A_j = \sum_{i=1}^j (n z_i)$  sont les totaux cumulatifs des  $n z_i$  dans

l'ordre en question, de sorte que  $0 = A_0 < A_1 < \dots < A_N = n$ . Soit  $u$  un nombre aléatoire uniforme dans l'intervalle  $[0, 1]$ . Les  $n$  unités qu'il faut inclure dans l'échantillon sont celles dont l'indice  $j$  satisfait  $A_{j-1} \leq u < A_j$  pour  $k = 0, 1, \dots, n-1$ . Soit  $s$  l'ensemble des  $n$  unités échantillonnées et  $\pi_i = P(i \in s)$  les probabilités d'inclusion de premier ordre. La méthode d'échantillonnage PPT systématique randomisée satisfait la condition

$$\pi_i = n z_i, \quad i = 1, 2, \dots, N. \quad (1.1)$$

Plusieurs autres méthodes d'échantillonnage sans remise qui satisfont (1.1) pour une taille d'échantillon fixe arbitraire  $n$  ont également été proposées dans la littérature, y compris la méthode d'échantillonnage avec probabilités inégales bien connue de Rao-Sampford (Rao 1965; Sampford 1967) et celles de Chao (1982), Chen, Dempster et Liu (1994), Tillé (1996), ainsi que Deville et Tillé (1998), entre autres. Les importants travaux de recherche sur les méthodes d'échantillonnage PPT ont été stimulés en grande partie par l'utilisation de l'estimateur d'Horvitz-Thompson (HT)  $\hat{T} = \sum_{i \in s} y_i / \pi_i$  pour le total de population  $T = \sum_{i=1}^N y_i$  d'une variable d'intérêt  $y$ . L'estimateur HT est très efficace quand  $y$  est fortement corrélé à la variable de taille  $x$  et que la méthode d'échantillonnage satisfait (1.1). Il s'agit de l'unique estimateur sans biais par rapport au plan parmi la classe d'estimateurs linéaires  $\sum_{i \in s} w_i y_i$  pour  $T$  si les poids  $w_i$  dépendent uniquement de  $i$ .

1. Mary E. Thompson, Département de statistique et de science actuarielle, Université de Waterloo, Courriel : methomps@uwaterloo.ca; Changbao Wu, Département de statistique et de science actuarielle, Université de Waterloo, Courriel : cbwu@uwaterloo.ca.

Zaslavsky, Zheng et Adams considèrent les taux d'échantillonnage optimaux dans des plans d'échantillonnage par élément, quand l'analyse prévue est la régression linéaire pondérée par les poids de sondage et que les quantités à estimer sont des combinaisons linéaires des coefficients de régression provenant d'un ou de plusieurs modèles. Ils commencent par élaborer des méthodes en supposant que des renseignements exacts sur les variables du plan existent dans la base de sondage, puis ils les généralisent à des situations où l'information pour certaines variables du plan n'est disponible que sous forme d'agrégat pour des groupes de sujets éventuels ou provient de données inexactes ou périmées. Les applications éventuelles comprennent l'estimation des moyennes pour plusieurs ensembles de domaines chevauchants, l'estimation pour des sous-populations telles que les races minoritaires par échantillonnage non proportionnel des régions géographiques, ainsi que les études où les caractéristiques pour lesquelles des données sont disponibles dans les bases de sondage sont mesurées avec une erreur.

L'article de Li porte sur le problème de l'estimation d'un total de population finie à l'aide d'un estimateur par la régression généralisée non linéaire. L'auteur utilise la méthode de transformation de Box-Cox ainsi que l'estimation par la méthode du pseudo maximum de vraisemblance pour obtenir des prédictions dictées par les données. Il montre que l'estimateur par la régression résultant est convergent sous le plan et évalue également les propriétés de cet estimateur au moyen d'une étude par simulation.

Bégin et Hulliger étendent l'algorithme BACON au traitement de données d'enquête incomplètes. L'algorithme BACON a été élaboré pour repérer les observations aberrantes multivariées en utilisant la distance de Mahalanobis. Quand des valeurs manquent, l'algorithme EM peut être utilisé pour estimer la matrice de covariance à chaque étape d'itération de l'algorithme BACON. Afin de pouvoir traiter l'échantillonnage en population finie, les auteurs modifient l'algorithme EM et obtiennent une version qu'ils nomment algorithme EBM (espérance estimée/maximisation) qu'ils combinent à l'algorithme BACON. Ils obtiennent ainsi l'algorithme BACON-EBM qu'ils appliquent à deux ensembles de données et comparant à d'autres méthodes.

L'article de Jäckle et Lynn offre une évaluation empirique des effets des paiements incitatifs continus sur l'attrition, le biais de non-réponse et la non-réponse partielle, ainsi que de la variation éventuelle de ces effets d'une vague à l'autre d'une enquête par panel à mode d'interview mixte réalisée auprès des jeunes au Royaume-Uni. Ils testent plusieurs hypothèses quant aux effets des primes d'encouragement. Ils concluent que l'offre de ces primes est un moyen efficace de maintenir la taille d'échantillon d'un panel, donc de s'assurer de sa valeur en ce qui concerne l'efficacité, surtout pour les analyses de sous-groupes. Cependant, ils constatent aussi que les primes d'encouragement n'ont aucun effet sur le biais dû à l'attrition.

Enfin, Mohadjer et Curtin discutent des défis associés à la conception et à la mise en œuvre d'un processus d'échantillonnage permettant d'atteindre les objectifs de la National Health and Nutrition Examination Survey (NHANES). Ils décrivent comment le plan d'échantillonnage de la NHANES doit établir un juste équilibre entre les exigences liées à l'obtention d'échantillons de sous-domaines efficaces, d'une part, et d'une charge de travail pouvant être bien gérée par le personnel chargé des interviews et des examens dans les centres d'examen mobiles (CEM), d'autre part, tout en maintenant les taux de réponse aussi élevés et les coûts faibles que possible. L'article décrit en détails plusieurs caractéristiques uniques du plan de sondage de la NHANES et se termine par un bref résumé des réalisations et de certains défis que posera dans l'avenir l'élaboration du plan de sondage de la NHANES.

Harold Mantel, Rédacteur en chef délégué

## Dans ce numéro

Ce numéro de *Techniques d'enquête* comprend des articles traitant de divers sujets de méthodologie.

Dans le premier article, Thompson et Wu considèrent le problème de l'obtention des probabilités d'inclusion pour le calcul des poids d'échantillonnage quand des modifications ou des compromis ont été apportés au plan d'échantillonnage original à cause de contraintes ou de limites pratiques. L'étude du problème a été motivée par l'enquête réalisée par la Chine dans le cadre du projet International Tobacco Control Policy Evaluation Survey en utilisant un plan d'échantillonnage à plusieurs degrés avec certains districts ont refusé de participer à l'enquête, des unités de substitution ont dû être sélectionnées après le tirage de l'échantillon original. Cette substitution a rendu difficile le calcul des probabilités d'inclusion de premier ordre et virtuellement impossible celui des probabilités de deuxième ordre. Dans l'article, les auteurs démontrent, théoriquement et empiriquement, que les probabilités d'inclusion de premier et de deuxième ordres peuvent être estimées avec précision au moyen de simulations Monte-Carlo. Torabi et Rao calculent l'erreur quadratique moyenne (EQM) d'un nouvel estimateur par la régression généralisée (GREG) d'une moyenne de petit domaine sous un modèle à deux niveaux et comparent, théoriquement et empiriquement, l'efficacité relative du nouvel estimateur GREG proposé et de l'estimateur fondé sur le meilleur prédicteur linéaire sans biais.

Dans son article, You discute de l'utilisation de divers modèles transversaux et chronologiques pour petits domaines en vue d'estimer le taux de chômage dans des régions infraprovinciales au Canada. En particulier, il considère un modèle non linéaire intégré à effets mixtes sous un cadre hiérarchique bayésien (HB). Il utilise une approche HB avec la méthode d'échantillonnage de Gibbs pour obtenir des estimations des moyennes et des variances à posteriori des taux de chômage dans les petits domaines. Le modèle HB proposé produit des estimations fondées sur le modèle fiables si l'on s'en tient à la réduction du coefficient de variation. You analyse aussi la qualité de l'ajustement du modèle proposé et compare les estimations fondées sur le modèle aux estimations directes.

L'article de Wang, Fuller et Qu traite de l'estimation pour petits domaines sous une contrainte. Les auteurs étudient l'effet de divers modèles augmentés en ce qui concerne l'EQM et le prédicteur EBLUP. Ils considèrent des modèles pour petits domaines augmentés d'une variable explicative supplémentaire pour lesquels les prédicteurs pour petits domaines ordinaires présentent la propriété d'autocontrôle. En outre, ils considèrent des modèles pour petits domaines augmentés d'une variable auxiliaire supplémentaire qui est fonction de la taille du domaine afin de réduire le biais lorsqu'un modèle incorrect est utilisé aux fins de prédictions.

Nandram et Choi exposent une approche intéressante de répartition des électeurs indécis dans les sondages réalisés avant une élection. Les données des sondages électoraux sont habituellement présentées dans des tableaux de contingence à double entrée et de nombreux sondages sont réalisés avant l'élection proprement dite. Les auteurs décrivent la construction et l'analyse d'un modèle de non-réponse non ignorable variant en fonction du temps par des méthodes bayésiennes. Ils comparent leurs modèles à des versions élargies (afin d'inclure la variable de temps) des modèles de non-réponse ignorable et non ignorable introduits par Nandram, Cox et Choi (*Techniques d'enquête*, 2005). Ils construisent aussi un nouveau paramètre pour faciliter la prédiction du gagnant. Ils illustrent leur approche à l'aide de données de sondage recueillies durant la campagne pour l'élection du gouverneur de l'Ohio de 1998.

Dans leur article, Lazar, Meeden et Nelson élaborent une approche bayésienne de l'échantillonnage d'une population finie qui s'appuie sur l'utilisation d'une loi a posteriori de Polya, quand l'information a priori est disponible sous la forme de données partielles sur une variable auxiliaire. Ils présentent l'estimateur de Polya contraint et montrent qu'il possède des similarités avec l'estimateur par la régression généralisée sous échantillonnage aléatoire simple. Toutefois, leur estimateur ne requiert pas la spécification d'un modèle linéaire. Il est également apparenté aux méthodes fondées sur la vraisemblance empirique. Des exemples servent à illustrer la théorie.





The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.



Le papier utilisé dans la présente publication répond aux exigences minimales de l'American National Standard for Information Sciences – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.

# Techniques d'enquête

## Une revue éditée par Statistique Canada

### Volume 34, numéro 1, juin 2008

#### Table des matières

Dans ce numéro.....1

#### Articles Réguliers

Mary E. Thompson et Changbao Wu	Échantillonnage PPT systématisé basé sur la simulation en cas de substitution d'unités.....3
Mahmoud Torabi et J.N.K. Rao	Estimation pour petits domaines sous un modèle à deux niveaux.....13
Yong You	Une approche intégrée de modélisation de l'estimation du taux de chômage pour les régions intraprovinciales au Canada.....21
Junyan Wang, Wayne A. Fuller et Yongming Qu	Estimation pour petits domaines sous une contrainte.....33
Balgobin Nandram et Jai Won Choi	Une répartition bayésienne des électeurs indécis.....41
Radu Lazar, Glen Meeden et David Nelson	Une approche bayésienne non informative de l'échantillonnage d'une population finie en utilisant des variables auxiliaires.....55

Alan M. Zaslavsky, Hui Zheng et John Adams	Répartition optimale de l'échantillon pour une régression convergente sous le plan dans une enquête sur les services aux cancéreux quand les variables du plan d'échantillonnage sont connues pour des agrégats.....71
--	--

Yan Li	Estimateurs par la régression généralisée d'un total de population finie utilisant la méthode de Box-Cox.....87
Cédric Béguin et Beat Hultiger	L'algorithme BACON-EEM pour la détection d'observations aberrantes multivariées dans des données d'enquête incomplètes.....99

Annette Jackle et Peter Lynn	Offre de primes d'encouragement aux répondants dans une enquête par panel multimodes : effets cumulatifs sur la non-réponse et le biais.....115
Leyla Mohadjer et Lester R. Curtin	Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey.....131

*Techniques d'enquête* est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

#### COMITÉ DE DIRECTION

##### Président

D. Royce

##### Anciens présidents

G.J. Brackstone  
R. Platak

#### COMITÉ DE RÉDACTION

##### Rédacteur en chef

J. Kovar, *Statistique Canada*  
H. Mantel, *Statistique Canada*

##### Rédacteurs associés

J.M. Brick, *Westat Inc.*

P. Cantwell, *U.S. Bureau of the Census*

J.L. Eltinge, *U.S. Bureau of Labor Statistics*

W.A. Fuller, *Iowa State University*

J. Gambino, *Statistique Canada*

M.A. Hidiroglou, *Statistique Canada*

D. Judkins, *Westat Inc.*

P. Kott, *National Agricultural Statistics Service*

P. Lahiri, *JPSM, University of Maryland*

P. Lavallée, *Statistique Canada*

G. Naitian, *Hebrew University*

J. Opsomer, *Colorado State University*

D. Pfeffermann, *Hebrew University*

N.G.N. Prasad, *University of Alberta*

J.N.K. Rao, *Carleton University*

##### Rédacteurs adjoints

J.-F. Beaumont, P. Dick, D. Haziza, Z. Patak, S. Rubin-Bleuer et W. Yung, *Statistique Canada*

#### POLITIQUE DE RÉDACTION

*Techniques d'enquête* publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généraux. Une importance particulière est accordée à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

#### Présentation de textes pour la revue

*Techniques d'enquête* est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à le faire parvenir en français ou en anglais en format électronique et préféablement en Word au rédacteur en chef, [rte@statcan.ca](mailto:rte@statcan.ca), Statistique Canada, 150 Promenade du Pré Tunney, Ottawa, (Ontario), Canada, K1A 0T6). Pour les instructions sur le format, veuillez consulter les directives présentées dans la revue.

#### Abonnement

Le prix de la version imprimée de *Techniques d'enquête* (N° 12-001-XPB au catalogue) est de 58 \$ CA par année. Le prix n'inclut pas les taxes de vente canadiennes. Des frais de livraison supplémentaires s'appliquent aux envois à l'extérieur du Canada. États-Unis 12 \$ CA (6 \$ x 2 exemplaires). Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research, la Société Statistique du Canada et l'Association des statisticiens et statisticiens du Québec. Des versions électroniques sont disponibles sur le site Internet de Statistique Canada : [www.statcan.ca](http://www.statcan.ca).



Ottawa

ISSN 0714-0045

Périodicité : semestrielle

N° 12-001-XPB au catalogue

Jun 2008

Tous droits réservés. Le produit ne peut être reproduit et/ou transmis à des personnes ou organisations à l'extérieur de l'organisme du détenteur de licence. Des droits raisonnables d'utilisation du contenu de ce produit sont accordés seulement à des fins de recherche personnelle, organisationnelle ou de politique gouvernementale ou à des fins éducatives. Cette permission comprend l'utilisation du contenu dans des analyses et dans la communication des résultats et conclusions de ces analyses, y compris la citation de quantités limitées de renseignements complémentaires extraits du produit. Cette documentation doit servir à des fins non commerciales seulement. Si c'est le cas, la source des données doit être citée comme suit : Source (ou « Adapté de », s'il y a lieu) : Statistique Canada, année de publication, nom du produit, numéro au catalogue, volume et numéro, période de référence et page(s). Autrement, les utilisateurs doivent d'abord demander la permission écrite aux Services d'octroi de licences, Division des services à la clientèle, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

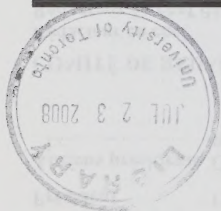
© Ministre de l'Industrie, 2008

Publication autorisée par le ministre responsable de Statistique Canada

Juin 2008 • Volume 34 • Numéro 1

# Une revue éditée par Statistique Canada

# Techniques d'enquête



## Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1-800-263-1136).

Pour toute demande de renseignements au sujet de ce produit ou sur l'ensemble des données et des services de Statistique Canada, visiter notre site Web à [www.statcan.ca](http://www.statcan.ca). Vous pouvez également communiquer avec nous par courriel à [infostats@statcan.ca](mailto:infostats@statcan.ca) ou par téléphone entre 8 h 30 et 16 h 30 du lundi au vendredi aux numéros suivants :

### Centre de contact national de Statistique Canada

Numéros sans frais (Canada et États-Unis) :

Service de renseignements	1-800-263-1136
Service national d'appareils de télécommunications pour les malentendants	1-800-363-7629
Télécopieur	1-877-287-4369

### Appels locaux ou internationaux :

Service de renseignements	1-613-951-8116
Télécopieur	1-613-951-0581

### Programme des services de dépôt

Service de renseignements	1-800-635-7943
Télécopieur	1-800-565-7757

## Comment accéder à ce produit ou le commander

Le produit n° 12-001-X au catalogue est disponible gratuitement sous format électronique. Pour obtenir un exemplaire, il suffit de visiter notre site Web à [www.statcan.ca](http://www.statcan.ca) et de choisir la rubrique « Publications » > « Publications Internet gratuites ».

Ce produit n° 12-001-X au catalogue est aussi disponible en version imprimée standard au prix de 30 \$CAN l'exemplaire et de 58 \$CAN pour un abonnement annuel.

Les frais de livraison supplémentaires suivants s'appliquent aux envois à l'extérieur du Canada :

	Exemplaire	Abonnement annuel
États-Unis	6 \$CAN	12 \$CAN
Autres pays	10 \$CAN	20 \$CAN

Les prix ne comprennent pas les taxes sur les ventes.

La version imprimée peut être commandée par les moyens suivants :

- Téléphone (Canada et États-Unis) 1-800-267-6677
  - Télécopieur (Canada et États-Unis) 1-877-287-4369
  - Courriel [infostats@statcan.ca](mailto:infostats@statcan.ca)
  - Poste
- Finances  
Statistique Canada  
Immeuble R.-H.-Coats, 6<sup>e</sup> étage  
150, promenade Tunney's Pasture  
Ottawa (Ontario) K1A 0T6

- En personne auprès des agents et librairies autorisés.

Lorsque vous signalez un changement d'adresse, veuillez nous fournir l'ancienne et la nouvelle adresse.

## Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois. À cet égard, notre organisme s'est doté de normes de service à la clientèle que les employés observent. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1-800-263-1136. Les normes de service sont aussi publiées sur le site [www.statcan.ca](http://www.statcan.ca) sous « À propos de nous » > « Offrir des services aux Canadiens ».



---

# Techniques d'enquête

---

N° 12-001-XPB au catalogue

Une revue

éditée

par Statistique Canada

Juin 2008

•  
Volume 34

•  
Numéro 1

